Ensemble Architectures and Fusion Techniques for Convolutional Neural Networks Applied to Medical Image Analysis

Cícero Lima Costa

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2024

Cícero Lima Costa

Ensemble Architectures and Fusion Techniques for Convolutional Neural Networks Applied to Medical Image Analysis

Tese de doutorado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Bruno Augusto Nassif Travençolo
Coorientador: Celia Aparecida Zorzo Barcelos

Uberlândia
2024

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

C837 Costa, Cícero Lima, 1989-
2024 Ensemble architectures and fusion techniques for convolutional neural networks applied to medical image analysis [recurso eletrônico] / Cícero Lima Costa. - 2024.

Orientador: Bruno Augusto Nassif Travençolo.
Coorientadora: Celia Aparecida Zorzo Barcelos.
Tese (Doutorado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.te.2024.618
Inclui bibliografia.
Inclui ilustrações.

1. Computação. I. Travençolo, Bruno Augusto Nassif, 1981-, (Orient.). II. Barcelos, Celia Aparecida Zorzo, 1954-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Tese, 29/2024, PPGCO				
Data:	26 de agosto de 2024	Hora de início:	08:00	Hora de encerramento:	11:35
Matrícula do Discente:	11923CCP007				
Nome do Discente:	Cícero Lima Costa				
Título do Trabalho:	Ensemble Architectures and Fusion Techniques for Convolutional Neural Networks Applied to Medical Image Analysis.				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	Da Visão Computacional à Visualização: novas proposições para pesquisa em eScience (Produtividade em Pesquisa - PQ - 306436/2022-1)				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Celia Aparecida Zorzo Barcelos FAMAT/UFU(Coorientadora), André Ricardo Backes - DC/UFSCar, Leandro Nogueira Couto - FACOM/UFU, João Fernando Mari - IEP/UFV, Vinicius Ruela Pereira Borges CIC/UnB e Bruno Augusto Nassif Travençolo - FACOM/UFU, orientador do candidato.

A Prof. Celia Aparecida Zorzo Barcelos - FAMAT/UFU, coorientadora do candidato, não pode participar por motivo de saúde.

Os examinadores participaram desde as seguintes localidades: João Fernando Mari - Rio Paranaiba/MG, André Ricardo Backes - São Carlos/SP. O aluno participou de Patrocínio-MG, os outros membros da banca participaram da cidade de Uberlândia.

Iniciando os trabalhos o presidente da mesa, Prof. Dr. Bruno Augusto Nassif Travençolo, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir ao candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Doutor.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **André Ricardo Backes**, **Usuário Externo**, em 26/08/2024, às 16:53, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Bruno Augusto Nassif Travençolo**, **Professor(a) do Magistério Superior**, em 26/08/2024, às 16:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Vinicius Ruela Pereira Borges**, **Usuário Externo**, em 26/08/2024, às 18:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **João Fernando Mari**, **Usuário Externo**, em 26/08/2024, às 21:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Leandro Nogueira Couto**, **Professor(a) do Magistério Superior**, em 27/08/2024, às 13:54, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5625037** e o código CRC **4A9C1F16**.

Referência: Processo nº 23117.053720/2024-41 SEI nº 5625037

I dedicate mainly to my wife, my parents and everyone who somehow contributed to the development of this work.

Acknowledgements

I thank...

God, for illuminating my way and allowing me to achieve this goal.

My wife, my parents, my sister and my nephews for always supporting and helping in difficult times.

The advisor Bruno Augusto Nassif Travençolo for his support and guidance at all times during this work.

The Co-supervisor Celia A. Zorzo Barcelos for her support.

My friends who collaborated in obtaining knowledge and proved to be companions during the course.

The IFTM for granting a license to carry out the work.

"The greatest discoveries often lie not in finding new things, but in seeing familiar things in new ways."
(Alexander Fleming)

Resumo

Algoritmos de visão computacional, como redes neurais convolucionais, são usados para automatizar processos na medicina e apoiar o diagnóstico. Esses algoritmos minimizam o erro humano durante a análise de imagens médicas e reduzem a variabilidade interoperador. Neste estudo, para apoiar o diagnóstico, foram propostas três estratégias envolvendo fusão de redes neurais convolucionais. Primeiro, comitês de redes neurais convolucionais foram utilizados na tarefa de classificação de imagens gastrointestinais. Em segundo lugar, através da fusão de modelos convolucionais, foi proposto um novo modelo para detectar pontos de referência em imagens de cefalogramas laterais, radiografias de mãos e radiografias de pulmão. A terceira análise testou se o pré-processamento de imagens ajudaria os modelos convolucionais na tarefa de detecção de pontos de referência e segmentação de regiões. As estratégias propostas foram avaliadas com base em métricas comuns na literatura, como erro radial médio e *F1-score*. Além disso, alinhado aos conceitos de computação verde, também foram avaliados o consumo de recursos e as emissões de poluentes. Para a tarefa de classificação, o comitê proposto obteve *F1-score* de $0,910$, correspondendo à literatura, porém, utilizando equipamentos de menor custo. Para detecção de pontos de referência, por meio de fusão de modelos, considerando a taxa de detecção de sucesso, *success detection rate* (SDR), entre os pontos de referência previstos e os pontos de referência originais, alcançamos SDR de $95,72\%$ para o cefalograma lateral e $99,56\%$ para as radiografias de mão, ambos considerando uma distância de até 4mm. Para radiografias de pulmão, obtivemos um SDR de $84,21\%$ considerando 6 pixels de distância. Nossa proposta também reduziu o tempo de execução, o consumo de energia e as emissões de carbono em cerca de 65%. A estratégia de pré-processamento não apresentou melhorias significativas nos resultados.

Palavras-chave: Aprendizagem profunda. Comitês. Fusão. Classificação. Raios X.

Abstract

Computer vision algorithms such as convolutional neural networks are used to automate processes in medicine and support diagnosis. These algorithms minimize human error during medical image analysis and reduces inter-operator variability. In this study, to support the diagnosis, three strategies involving fusion of convolutional neural networks were proposed. First, ensemble architectures were used in the gastrointestinal image classification task. Second, through the fusion of convolutional models, a new model was proposed to detect landmarks in images of lateral cephalograms, hand X-rays and lung X-rays. The third analysis tested whether image preprocessing would help convolutional models in the task of landmark detection and region segmentation. The proposed strategies were evaluated based on common metrics in the literature such as mean radial error and F1-score. In addition, aligning with the concepts of green computing, resource consumption and pollutant emissions were also evaluated. For the classification task, the proposed ensemble achieved an F1-score of 0.910, matching the literature, however, using lower cost equipment. For landmark detection, through model fusion, considering the success detection rate (SDR) between the predicted landmarks and the original landmarks, we achieved SDR of 95.72% for the lateral cephalogram and 99.56% for the hand x-rays, both considering a distance up to 4mm. For lung x-rays, we obtained an SDR 84.21% considering 6 pixels of distance. Our proposal also reduced execution time, energy consumption and carbon emissions by around 65%. The preprocessing strategy showed no with significant improvements over the results.

Keywords: Deep learning. Ensamble. Fusion. Classification. X-rays.

List of Figures

List of Tables

Acronyms list

AI Artificial Intelligence

ANN Artificial Neural Network

AFPF Attentive Feature Pyramid Fusion

CNN Convolutional Neural Networks

DNN Deep Neural Networks

DL Deep Learning

GPU Graphic Processing Unit

GI Gastrointestinal

HD Hausdorff distance

LR Learning Rate

MCC Matthews correlation coefficient

MRE Mean Radial Error

RNN Recurrent Neural Networks

SbP Similar based-Paper

SGD Stochastic Gradient Descent

SDR Sucess Detection Rate

YOLO You Only Look Once

Contents

Chapter 1

Introduction

Image analysis is widely used in medicine for diagnosis, surgical procedures, and various other tasks. However, one of the challenges in the medical field is inter-operator variability. Different specialists analyzing the same exam at different times can produce different reports. To minimize this problem, computational resources can be utilized (ZENG et al., 2021; RUNDO et al., 2020; LITJENS et al., 2017).

Improving and interpreting images using computational resources are goals of the computer vision field. This area has developed and refined algorithms over many years of research, ranging from traditional techniques such as preprocessing, filtering, and segmentation to recent advances provided by Deep Neural Networks (DNN), a subtype of Artificial Neural Network (ANN). While the typical structure of an ANN consists of only a few layers, DNNs have a deeper architecture with more hidden layers, making them more effective in solving complex problems (STOCKMAN; SHAPIRO, 2001; BAYOUDH et al., 2021; RUNDO et al., 2020).

The most common type of DNN for image processing is the Convolutional Neural Networks (CNN). CNNs can be used to classify and detect regions of interest in medical images (SANTOS et al., 2021; KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SZE et al., 2017).

Currently, there are several types of CNN architectures with varying numbers of layers, designed for different types of problems. The expansion of CNNs has been driven by computational advances of the 21st century and, particularly after 2016, by the use of graphics processing units Graphic Processing Unit (GPU). Because of their high parallel processing capacity, GPUs allow the training of deep CNNs in a relatively short time (SZE et al., 2020; LECUN; BENGIO; HINTON, 2015).

The diversity of CNNs creates opportunities for researchers to integrate different CNN architectures into their analyses. For instance, some studies propose using CNN ensembles to solve problems. In an ensemble, several CNNs are combined before the system provides the final answer. Other studies propose the fusion of networks at different levels of their architecture (BAYOUDH et al., 2021; RADEVSKI; BENNANI, 2000).

In this work we will explore the use of an ensemble of CNNs and fusion alternatives in CNN architectures, in order to contribute to computational solutions to support analysis of medical images. Furthermore, the resource consumption of the models studied will be analyzed to indicate solutions that consume less resources.

1.1 Motivation

In the analysis of medical images using computer vision algorithms it is crucial to minimize errors as much as possible – although some situations allow for a margin of error. There are instances where computer vision algorithms, such as CNNs, achieve nearly 100% accuracy on specific datasets. However, there is always room for improvement in other contexts. Additionally, improvements can be made in terms of resource consumption by proposing models that achieve objectives more quickly and with less energy use.

In this work, we analyzed the performance of CNNs for two tasks: (i) classification of gastrointestinal images and (ii) detection of landmarks in images of lateral cephalograms, hand radiographs, and lung radiographs.

The first part of the study focuses on classifying images of gastrointestinal diseases using CNNs, justified by the significant negative impact these diseases have on the global population. In (BORGLI et al., 2020), the authors present a dataset of gastrointestinal images and report that gastrointestinal cancer results in about 3.5 million new cases per year worldwide and has a high mortality rate. Using computational resources to aid diagnosis allows for earlier cancer detection and may reduce the cost of exams.

The other part of the study focuses on detecting landmarks in medical images using CNNs, which is important because physicians often need to analyze specific regions in a medical image. For example, analyzing lung regions to diagnose respiratory diseases such as COVID-19, which affected many people worldwide and requires accurate, quick, and cost-effective diagnosis (GIEŁCZYK et al., 2022; GAGGION et al., 2022).

Another example of the importance of landmark detection is in lateral cephalograms. In these images, orthodontists need to identify landmarks to diagnose a patient's craniofacial condition and plan treatment (ZENG et al., 2021).

CNNs are already being used for both image classification and landmark detection in medical images. However, there is still potential for improvement, as discussed in (MESQUITA et al., 2023). In works such as (BORGLI et al., 2020) and (BAYOUDH et al., 2021), the authors discuss enhancing CNN performance through strategies such as CNN ensembles and fusions, which are the strategies investigated in this work.

1.1.1 Goals

The main goal of this work is to investigate the use of ensemble classifiers and the influence of fusion methods on the performance of CNNs in medical imaging. The specific

goals are:

- ❏ Investigate CNN models used for image classification. Explore various convolutional neural network architectures and their applications in image classification, evaluating the models' performance in terms of accuracy, efficiency, and generalization.
- ❏ Investigate CNN models used for landmark detection. Examine CNN architectures specifically designed to detect landmarks in images, assessing their performance in identifying landmarks and their reliability across different medical image datasets.
- ❏ Develop and analyze an ensemble architecture for image classification. Design and implement an ensemble approach that combines multiple CNN models for image classification. Analyze the performance gains and potential benefits of using ensembles compared to single CNN models.
- ❏ Compare the performance of CNNs using different types of fusion. Investigate various data fusion techniques (such as early fusion, late fusion, and hybrid fusion) and their impact on CNN performance. Conduct experiments to understand how these fusion methods affect the accuracy and robustness of CNN models.
- ❏ Explore image preprocessing methods. Analyze if these preprocessing techniques, when combined with fusion methods, can improve the performance of CNNs in landmark detection.
- ❏ Contribute to the improvement of landmarks detection techniques. Develop new methodologies or refine existing techniques to increase the accuracy and reliability of landmark detection, sharing discoveries and insights that can advance the field and benefit other researchers and professionals.
- ❏ Develop and analyze a universal CNN model for landmark detection. Create a versatile CNN model capable of detecting landmarks in different datasets.
- ❏ Propose solutions aligned with green computing. Explain the cost and resource consumption for the analyzed models and the proposed models, aiming for more efficient and environmentally friendly computing solutions.

1.2 Hypothesis

CNNs are a great advance for the image classification task, but there is still possibility of improvement. This work is being developed on the hypothesis that CNN ensembles and fusion strategies can generate better results in classification tasks and landmark detection.

Questions to be answered:

1. Does the use of ensemble classifiers improve the result of CNNs in classification tasks?

2. Does the use of fusion strategies in CNN architectures improves the performance in landmark detection tasks?

3. Does the use of preprocessing strategies improves the performance of CNNs applied to landmark detection?

1.3 Contributions

1. Comparison of performance of CNNs in image classification task.

2. Creation of alternative CNN ensembles for image classification.

3. Performance comparison of CNN models for landmark detection.

4. Proposal of a method for image preprocessing and evaluation of whether preprocessed images enhance CNN performance in landmark detection and region segmentation tasks.

5. Development of a CNN model based on fusion strategies for landmark detection.

1.4 Thesis Organization

This work is organized into an introduction (1), a chapter on theoretical foundations (2), three chapters (3, 4, 5) detailing the proposals and their respective background, results, and conclusions, and finally a chapter summarizing the overall conclusions (6).

The theoretical foundation, presented in Chapter 2, reviews some basic concepts about digital imaging, computer vision, deep learning, ensembles and CNN fusion.

In Chapter 3, a proposal related to the use of ensembles to improve the results of a classification task is presented. This chapter presents related work, experiments, results and conclusions on the use of ensemble for classifying images of the gastrointestinal tract.

In Chapter 4, a universal CNN model for landmark detection is described. This chapter proposes fusion alternatives between models used in landmark detection. Experiments and results are presented based on the original models and the proposed fusion-based models. Finally, conclusions are presented based on the performance of distance metrics and the consumption of computational resources.

In Chapter 5, we analyze whether image preprocessing brings gains to CNN models. This chapter proposes a preprocessing approach, presents experimental results, and draws conclusions based on the findings.

Finally, Chapter 6 summarizes the conclusions drawn from this thesis, highlights achievements, discusses challenges encountered, and suggests directions for future research.

Chapter 2

Theoretical foundation

Technological advances occurring in the 21st century allow for an increase in the storage and dissemination of information, in text, image and video formats. Image processing is important in areas such as medical imaging, autonomous vehicles, facial recognition and industrial automation. Analyzing the content of images is a task in the area of computer vision. Computer vision systems can automate complex tasks, improve efficiency, and enable innovative solutions across multiple domains. CNNs are computer vision systems that are highly effective in tasks such as image classification, object detection and facial recognition due to their ability to capture spatial hierarchies and patterns in images (STOCKMAN; SHAPIRO, 2001; MODERSITZKI, 2009; AGRAWAL et al., 2011; WOODS, 2011; VOULODIMOS et al., 2018; BAWDEN; ROBINSON, 2020).

CNNs can operate individually or as ensembles. An ensemble of classifiers combines multiple learning algorithms to achieve superior predictive performance compared to individual models. This approach harnesses the strengths of different models to mitigate individual weaknesses, thereby aiming for more robust and accurate predictions (KRAWCZYK; SCHAEFER, 2014; KIM et al., 2015).

In addition to ensembles, CNN models can be combined in various ways, a concept known as CNN fusion. CNN fusion involves integrating multiple convolutional neural networks to create a more comprehensive and accurate model. This process may include combining different CNN architectures, layers, or features to leverage the strengths of each component. Fusion aims to enhance the model's ability to generalize across diverse datasets and conditions, making it invaluable for advanced computer vision tasks (RUNDO et al., 2020; BAYOUDH et al., 2021; HÖHN et al., 2021; GAGGION et al., 2021; GAGGION et al., 2022).

Lastly, it is essential to consider the resource consumption associated with computer vision tools, aligning with efforts in green computing. This field strives to reduce energy consumption and minimize the ecological footprint of computing activities. As demand for computing power grows, especially with the surge in deep learning and AI, green computing becomes increasingly critical. By optimizing algorithms, hardware, and

data centers for energy efficiency, green computing helps mitigate the environmental impact of technology, promotes sustainability, and ensures responsible use of computing resources (STRUBELL; GANESH; MCCALLUM, 2019; ANTHONY; KANDING; SELVAN, 2020; BENDER et al., 2021; SELVAN et al., 2022; MASLEJ et al., 2024).

This chapter provides an insight into computer vision, convolutional neural networks, ensemble of classifiers, data fusion, and green computing.

2.1 Computer vision

Research in computer vision drives the development and enhancement of computational techniques that enable computers to detect and locate objects in images and videos.

Image classification involves identifying and categorizing the content of images. For instance, in medical imaging, a physician can classify gastrointestinal images as either disease-free or showing conditions like polyps or gastritis. Landmark detection and segmentation are techniques used to pinpoint specific features and accurately delineate different parts of an image. Point detection entails identifying and locating specific points of interest within an image, such as anatomical landmarks critical for monitoring dysfunction progression or surgical planning. Segmentation divides an image into segments corresponding to distinct regions or objects within the image (MURPHY et al., 2006; JABRI et al., 2000; PARAGIOS; TZIRITAS, 1999; SWENSSON, 1996).

Since the inception of computer vision research, various techniques have been devised to enhance computational performance in object classification and detection tasks. Methods such as filters, image enhancement, and machine learning have been pivotal. As of 2014, CNN has shown promising results for object detection and localization in images. (GIRSHICK et al., 2014; GIRSHICK et al., 2015). According to Figure 1, deep learning is a subdomain of machine learning that is part of the field of Artificial Intelligence (AI) and is based on the functioning of the brain (SZE et al., 2017; LECUN; BENGIO; HINTON, 2015; RUNDO et al., 2020; SZE et al., 2020).

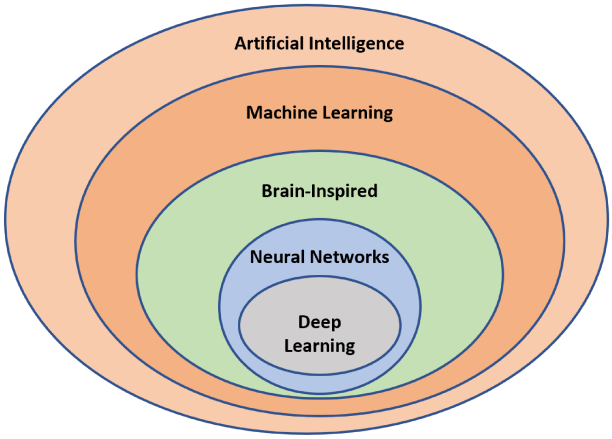


Figure 1 – Fields of Artificial Intelligence. Adapted from (SZE et al., 2020).

Various CNN architectures have shown promising results across different applications; however, opportunities for improvement remain. In more complex scenarios, combining CNNs with different architectures in an ensemble approach has been beneficial. Ensembles aggregate responses from classifiers with diverse characteristics and competencies to optimize classification accuracy (LECUN; BENGIO; HINTON, 2015; KIM et al., 2015; SZE et al., 2017; JO; NHO; SAYKIN, 2019; RUNDO et al., 2020; SZE et al., 2020).

Another strategy to enhance CNN performance involves integrating secondary information. Models fusion techniques facilitate the incorporation of supplementary data. Fusion can occur at different stages: early fusion integrates data before CNN processing, late fusion combines data post-processing, and intermediate fusion merges characteristics of both approaches (GRAPOV et al., 2018; RUNDO et al., 2020; BAYOUDH et al., 2021; HÖHN et al., 2021).

2.2 Digital image

A digital image is derived from an analog signal through processes of sampling and quantization. The digital image is a matrix formed by ϕ pixels, such that $\phi = (x, y)$, and the intensity of ϕ, given by $f(\phi)$, according to Equation 1. The digital image is represented as a matrix $N \times M$ (WOODS, 2011; GONZALEZ; WOODS, 2009; MODERSITZKI, 2009; GOSHTASBY, 2005)

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,M-1) \\ f(1,0) & f(1,1) & \dots & f(1,M-1) \\ \vdots & & & \\ f(N-1,0) & f(N-1,0) & \dots & f(N-1,M-1) \end{bmatrix} \tag{1}$$

2.3 Convolutional neural network models

Since the emergence of computers, there have been researches with the objective of making the computer reproduce biological characteristics; these are bioinspired researches. Among the bioinspired researches, there is one that simulates the functioning of the brain through artificial neural networks. The networks undergo many transformations, with proposals for different architectures to work with different types of data, as reported in the works of (FUKUSHIMA, 1980; LECUN et al., 1989; KRIZHEVSKY; SUTSKEVER; HINTON, 2012; LECUN; BENGIO; HINTON, 2015; SZEGEDY et al., 2017; SZE et al., 2017). Among the types of networks proposed throughout history, in this work, the focus is to use CNN.

After 2006, with the adoption of data processing using GPU, it became possible to train networks faster. The gain in training time leads to an increase in the number of layers in the networks, and consequently, there is an improvement in performance. In the 2012 ImageNet competition, CNN far outperforms the other competitors, nearly halving the error rates of the best competitors. In Figure 2, it is possible to observe the performance obtained by networks of different architectures between the years 2012 and 2015; it is noted that the performance in 2015 was better than in 2012, the error decreased, and the number of layers of the networks increased (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; LECUN; BENGIO; HINTON, 2015; RUSSAKOVSKY et al., 2015; SZE et al., 2017; SZE et al., 2020).

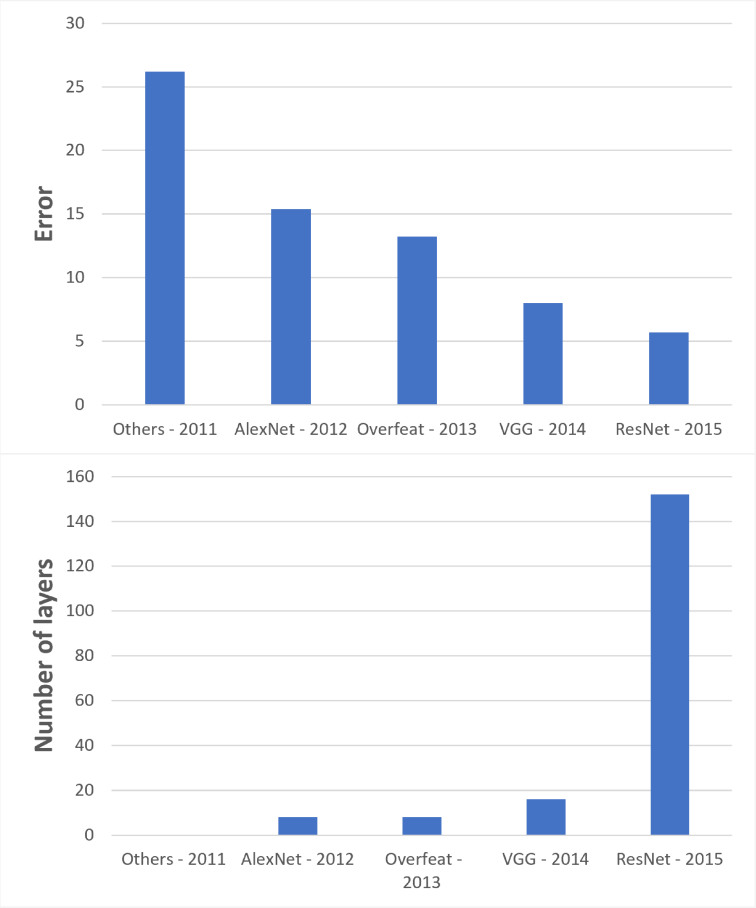


Figure 2 – Evolution of the performance of CNNs in the ImageNet competition. From top to bottom: the first graph shows that error decreased between 2011 and 2015; over the same period, the second graph shows that the number of layers in neural networks has increased. Adapted from (SZE et al., 2020) and (RUSSAKOVSKY et al., 2015)

The fundamental architecture of a CNN comprises an input layer, convolutional layers, fully connected layers, and an output layer. Additionally, pooling layers, activation functions, softmax function, and dropout are important in CNN performance. (LECUN; BENGIO; HINTON, 2015; SZE et al., 2017; KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Figure 3 shows the architecture of the AlexNet network, the winner of the ImageNet competition in 2012 (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Among some convolution layers, there is pooling layers, which is a type of subsampling that can take into account the mean, sum, or maximum value (max pooling) in a sample. An activation function is used between the convolution layers, and such a function can be a sigmoid, hyperbolic tangent, ReLu, Leaky ReLu, or exponential LU, with the ReLu function, $f(x) = \max(0, x)$, being the most traditional in CNNs. After the fully connected layer, the softmax function is used, establishes the percentage of probability of the input object about the classes of the problem. During the training of the network, dropout is

used, which forces neurons to learn to value communication with other neurons and not just a fixed flow of communication, preventing the network from specializing just to the training dataset (LECUN; BENGIO; HINTON, 2015; SZE et al., 2017; KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

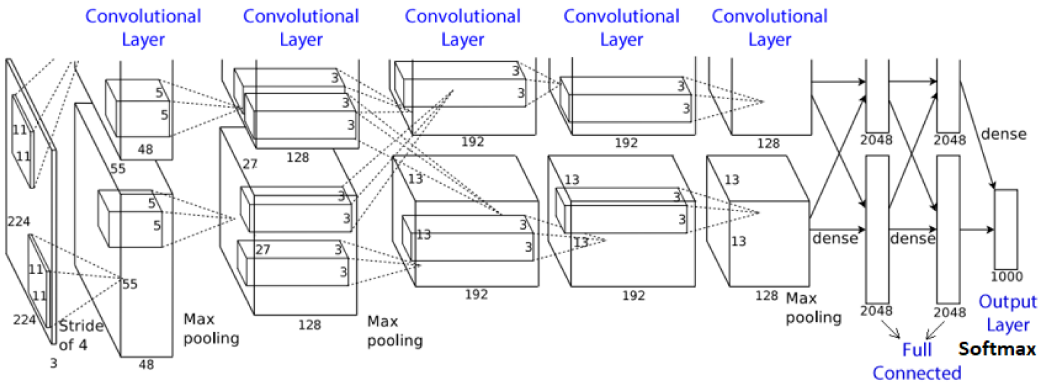


Figure 3 – AlexNet network architecture. Adapted from: (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Since AlexNet, numerous CNN architectures have been proposed. These architectures have significantly advanced object detection and classification in images, each offering unique features and achieving notable results in diverse contexts.

The VGG algorithm is a popular CNN architecture known for its simplicity and effectiveness. VGG is still popular for medical data (LITJENS et al., 2017), though recent landmark studies are using VGG16 (TAMMINA, 2019). It consists of 16 convolutional layers followed by fully connected layers. The VGG16 algorithm employs small convolutional filters and max-pooling layers in a homogeneous architecture. Despite its larger parameter count, VGG16 has demonstrated strong performance in various computer vision tasks, particularly image classification (LITJENS et al., 2017).

DenseNets are Densely Connected deep neural networks that tackle the vanishing gradient problem and promote feature reuse (HUANG et al., 2017). They introduce a dense connectivity pattern where each layer is directly connected to every other layer. This facilitates information flow and encourages feature reuse, leading to improved gradient flow and enhanced model performance. There are variations of the DenseNet architectures, such as DenseNet121, DenseNet161, and DenseNet201. The numbers $(121, 161, 201)$ represent the number of layers in each respective DenseNet variant. On the other hand, EfficientNet is a family of CNNs that achieve state of the art performance with computational efficiency (TAN; LE, 2019). These networks employ a compound scaling method that uniformly scales the network width, depth, and resolution. By optimizing the scaling coefficients, EfficientNet achieves high accuracy while maintaining a smaller model size and reduced computational requirements (TAN; LE, 2019). Efficientnet_b0 is a spe-

cific variant of the EfficientNet architecture, where "b0" indicates the baseline variant of EfficientNet.

The class of algorithms called MobileNet is a lightweight CNN designed for mobile and embedded vision applications. It utilizes depth-wise separable convolutions, splitting the standard convolution operation into depth-wise and point-wise convolutions. MobileNetV2 (SANDLER et al., 2018) improves the state of the art performance of mobile models on multiple tasks and benchmarks as well as across a spectrum of different model sizes (HOWARD et al., 2017). This reduces parameters and computations, making MobileNetV2 models suitable for resource-constrained environments, such as images or videos (DONG et al., 2020), without compromising performance.

ResNet introduced residual connections, enabling the network to learn residual mappings and overcome the vanishing gradient problem (LITJENS et al., 2017; HE et al., 2016). ResNet architectures, like ResNet-152, have achieved exceptional performance in image classification tasks, revolutionizing deep learning (BORGLI et al., 2020; HE et al., 2016).

In addition to the use of different architectures or different types of networks, there may also be the combination of networks in ensembles; in (LECUN; BENGIO; HINTON, 2015), the authors present the possibility of combining CNNs with Recurrent Neural Networks (RNN) that use reinforcement learning to decide which region of the image to look at.

Allied to deep networks, to improve performance in object classification and detection tasks, it is possible to combine resources through fusion, a subject that is detailed in the next section.

2.4 Ensemble of Classifiers

Collaborative decision-making involving individuals with diverse characteristics is commonplace in human interactions, particularly in business environments. This behavior can be replicated by computers through ensemble classifiers. Ensembles consistently yield strong results in classification tasks; however, their performance heavily relies on selecting appropriate classifiers and effective methods for combining classifier responses. In an ensemble, classifiers with varying characteristics and capabilities collaborate to derive optimal solutions for classification problems, as depicted in Figure 4 (RADEVSKI; BENNANI, 2000; AKSELA, 2003; KIM et al., 2015).

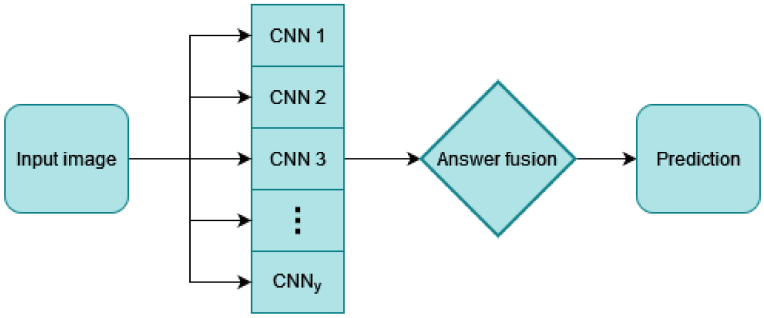


Figure 4 – Example architecture for ensemble of classifiers.

In ensembles architectures of classifiers, generating the final answer is a crucial task, and it is necessary to choose a good method (fusion method) for collective decision-making; methods based on static weights, voting, simple average, or weighted exponential average are often used (KRAWCZYK; SCHAEFER, 2014; KIM et al., 2015). In (KRAWCZYK; SCHAEFER, 2014), instead of traditional fusing methods, a perceptron neural network was used to fuse the ensemble response, which increased the overall classification accuracy.

The use of an ensemble of classifiers is reported in several works, such as (KRAWCZYK; SCHAEFER, 2014), in which the authors propose a less aggressive and more efficient approach for the diagnosis of breast cancer, performing the identification of cancer through thermal images. Additionally, an ensemble of classifiers is used to automatically classify features extracted from breast thermography. In (KIM et al., 2015), the authors use a database with images of faces from the real world, Static Facial Expressions in the Wild (SFEW 2.0), and they propose an ensemble of convolutional neural networks (CNNs) to analyze facial expressions. The returns generated by the networks are pooled based on the exponentially weighted average and simple average rule; the class that was chosen by more classifiers is defined as the ensemble's answer.

2.5 Convolutional Neural Networks Fusion

A physician, when diagnosing a patient, can rely on information from imaging tests, blood tests, and also on information arising from questions asked to the patient. The volume of patient data has become greater due to the emergence of new resources for acquiring them. This increase in data overwhelms physicians during decision-making. To aid in medical diagnosis, the use of algorithms to integrate and process large volumes of data is increasing.

Relevant information can be learned by algorithms automatically from images. Heterogeneous and multimodal imaging data can be integrated (undergo fusion) to reduce randomness and redundancy in order to improve diagnosis (RUNDO et al., 2020; BAYOUDH et al., 2021; HÖHN et al., 2021).

There are cases in which a fusion occurs between features generated by different CNNs trained on the same dataset. In other situations, features from CNNs of the same type, trained on different parts of the dataset, are fused (RUNDO et al., 2020; BAYOUDH et al., 2021; HÖHN et al., 2021).

Different types of images can be fused and made available to a classifier; the fusion can be early, in which the data will be integrated before being made available to the classifier, as shown in Figure 5. There is late fusion, in which the different types of data are passed to different classifiers, and the responses of the classifiers are integrated, as shown in Figure 6. Finally, there is the intermediate fusion, which joins the concepts of early fusion and late fusion, as shown in Figure 7 (BAYOUDH et al., 2021).

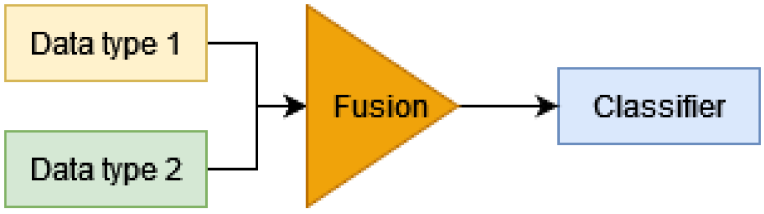


Figure 5 – Early Fusion. Adapted from: (BAYOUDH et al., 2021).

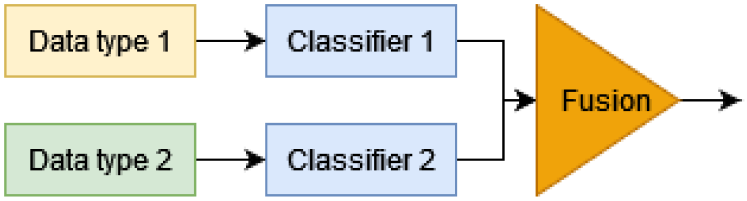


Figure 6 – Late fusion. Adapted from: (BAYOUDH et al., 2021).

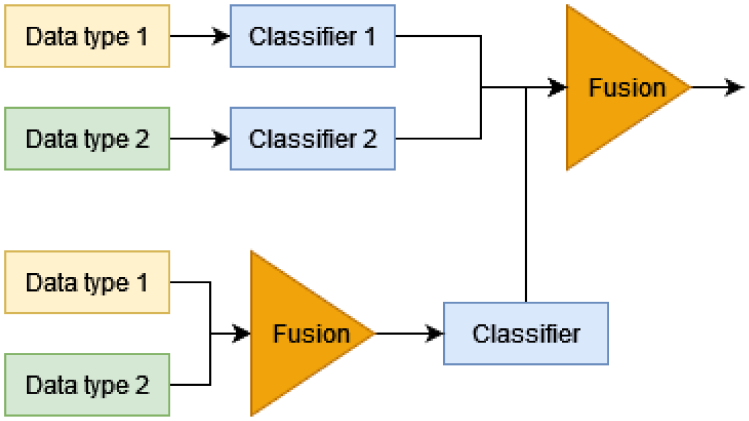


Figure 7 – Intermediate fusion. Adapted from: (BAYOUDH et al., 2021).

In literature we can find examples of the use of data fusion to improve the classification task. In (LI et al., 2020a), the authors propose the fusion of dermoscopic images with

other patient data for the diagnosis of skin diseases. The authors emphasize that there are few studies focused on data fusion for the classification of medical images (LI et al., 2020a).

In another work, the authors in (VADDI; MANOHARAN, 2020) propose the fusion of spectral and spatial information for the classification task using a simple design CNN network architecture. In (GRAPOV et al., 2018) emphasizes the possibility of integrating omic data with other types of data, including medical images. In (ADERGHAL et al., 2017) explore the possibility of fusing information from different brain projections using a CNN architecture. Instead of working with the entire volume of the brain, the authors propose the fusion of CNNs from the sagittal, coronal, and axial projections of the brain.

Based on related work, it can be seen that data fusion is a resource explored as a way to improve classification tasks.

2.6 Green computing: energy consumption and carbon dioxide emissions in computational applications

The adoption of GPU-based data processing has facilitated the proliferation of CNN models with increased layers, resulting in substantial gains in accuracy. However, this trend has also led to heightened GPU usage, thereby increasing energy consumption and carbon dioxide emissions, contributing to global warming (STRUBELL; GANESH; MCCALLUM, 2019; ANTHONY; KANDING; SELVAN, 2020; BENDER et al., 2021; SELVAN et al., 2022; MASLEJ et al., 2024).

In (HODAK; GORKOVENKO; DHOLAKIA, 2019), experiments were conducted on Lenovo ThinkSystem SR670 servers, primarily utilized for Deep Learning (DL). The findings revealed that GPUs accounted for 70% of energy consumption during the execution of CNN models, while the CPU and RAM accounted for 15% and 10%, respectively.

In order to align computing practices with global efforts to mitigate carbon emissions, developers must monitor the energy consumption and carbon emissions associated with training CNN models (ANTHONY; KANDING; SELVAN, 2020; HENDERSON et al., 2020; SELVAN et al., 2022). Several tools have been proposed and made available to calculate energy consumption and carbon emissions during CNN training, such as those presented in (LACOSTE et al., 2019; HENDERSON et al., 2020; ANTHONY; KANDING; SELVAN, 2020; BUDENNYY et al., 2022).

In this work, we will use the tool proposed by (ANTHONY; KANDING; SELVAN, 2020)[1] to monitor the energy consumption and carbon emissions associated with the studied models. This tool enables consumption predictions from the initial training epoch,

[1] <https://github.com/lfwa/carbontracker>

supports various environments, and automatically reports the average gCO_2/kWh for the region in which the server operates.

Chapter 3

Ensemble Architectures Analysis in Classification Task

The human Gastrointestinal (GI) tract is susceptible to several abnormal mucosal findings, including life-threatening diseases (BORGLI et al., 2020). GI cancer alone accounts for millions of new cases annually, emphasizing the need for improved endoscopic performance and systematic screening (JHA et al., 2020). Gastrointestinal exams and colonoscopy are essential procedures to investigate the human GI tract (JHA et al., 2021). These tests play a vital role in the diagnosis and management of gastrointestinal conditions, contributing to the early detection, treatment and prevention of serious complications (HICKS et al., 2021; JHA et al., 2021; BORGLI et al., 2020). However, current endoscopic scoring systems lack standardization and are subjective (JHA et al., 2021; BORGLI et al., 2020).

In this context, artificial intelligence (AI) enabled computer-assisted diagnostic systems, particularly machine learning, have shown promise in healthcare, but the scarcity of medical data impedes progress (BORGLI et al., 2020; JHA et al., 2020). To solve this, we used a dataset, called HyperKvasir, a large dataset of gastrointestinal images and videos collected during real exams (BORGLI et al., 2020). The dataset contains over 1.1×10^5 images and 374 videos and representing anatomical landmarks as well as pathological and normal findings (BORGLI et al., 2020).

Over the years, machine learning has evolved into deep learning algorithms, relying primarily on the DNN. Convolutional neural networks (CNN), a type of DNN, have emerged as a powerful tool for image analysis and classification, including medical imaging tasks. CNN ensemble architectures have been widely employed to improve predictive accuracy by combining the outputs of various models. These sets leverage the diversity of individual CNN models to improve overall performance. In addition, fusion techniques are employed to effectively integrate predictions from multiple CNN models (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SZE et al., 2017).

In this work, based on (COSTA et al., 2023), our main objective is to propose a new

ensemble architecture and efficient fusion techniques for CNNs in the classification of GI tract diseases using the HyperKvasir dataset, aiming to obtain better results than in the literature and to optimize computational resources. To achieve this, we performed a thorough literature review to identify relevant studies on the use of deep learning methods in similar health domains. In addition, we performed several experiments to evaluate the effectiveness of our proposed approach.

3.1 Background

Since the emergence of computers, there have been research efforts to make them reproduce biological characteristics; these are known as bioinspired research. Among the bioinspired research, there is one that seeks to simulate the functioning of the brain through artificial neural networks. These networks have undergone many transformations, as reported in the papers (FUKUSHIMA, 1980; LECUN et al., 1989; KRIZHEVSKY; SUTSKEVER; HINTON, 2012; LECUN; BENGIO; HINTON, 2015; SZEGEDY et al., 2017; SZE et al., 2017; SZE et al., 2020). This section provides an overview of the HyperKvasir database (BORGLI et al., 2020), the dataset utilized in this study. We reviewed the literature on deep learning in digital imaging and consider general model (BORGLI et al., 2020) as a reference for our research. Our objective is to establish a robust foundation by analyzing the dataset and surveying related studies.

3.1.1 HyperKvasir Dataset

The HyperKvasir dataset[1] is composed of images and videos. The dataset content was collected between 2008 and 2016, in a hospital in Norway. In this work, 10639 instances available in the dataset are used. The images are separated into 23 classes, in Figure 9 it is possible to see examples of images contained in the dataset. The Classes are Z-line, Pylorus, Retroflex stomach, Barrett's, Short segment, Esophagitis grade A, Esophagitis grade B-D, Cecum, Retroflex rectum, Terminal ileum, Polyps, Ulcerative colitis grade 0-1, Ulcerative colitis grade 1, Ulcerative colitis grade $1-2$, Ulcerative colitis grade 2, Ulcerative colitis grade $2-3$, Ulcerative colitis grade 3, Hemorrhoids, Dyed lifted polyps, Dyed resection margins, BBPS $0-1$, BBPS $2-3$, Impacted stool. The dataset offers a file (`.csv`) with the division of classes studied by Borgli (BORGLI et al., 2020).

3.1.2 Related Works

In our study, we started by establishing a solid foundation using the reference article (BORGLI et al., 2020). Expanding upon this work, we created a comprehensive graph, illustrated in Figure 10, to visually illustrate the interconnectedness of relevant papers

[1] Available at: <https://datasets.simula.no/hyper-kvasir>.

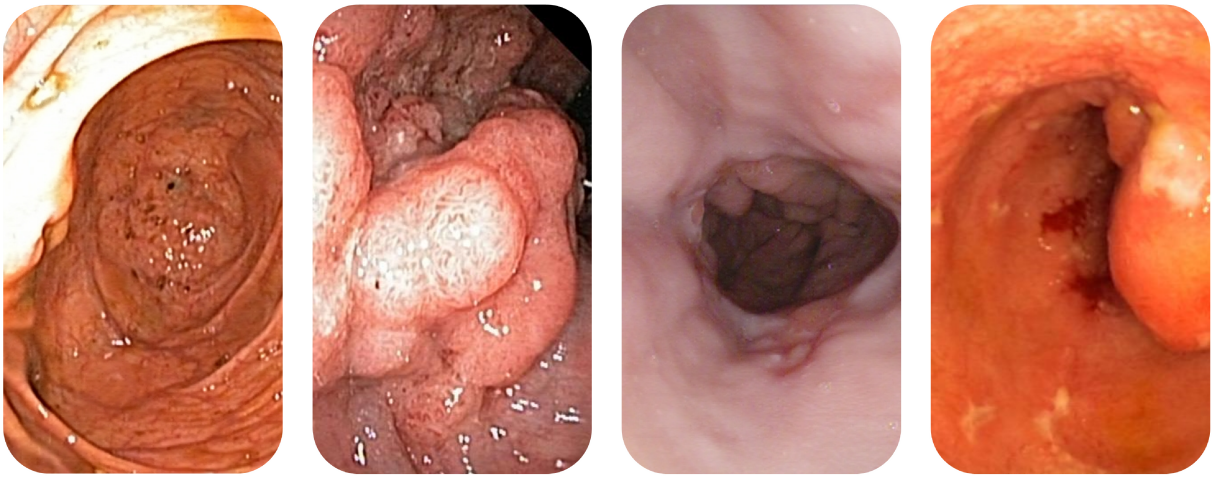


Figure 8 – Example of images present in the gastrointestinal disease image database. From left to right: image of cecum class, polyp class, esophagitis grade-a class, polyp class.

in our research field. To create the graph we used the online tool Connected Papers[2], the metrics of this tool are based on Co-citation and Bibliographic Coupling (KESSLER, 1963; OSAREH, 1996). This graph provides a valuable representation of the network of related literature, with a specific focus on the HyperKvasir image and video dataset for gastrointestinal endoscopy, as discussed by (BORGLI et al., 2020).

Upon analyzing the graph, we identified a total of 41 studies connected to the article (BORGLI et al., 2020), resulting in a set of 42 relevant studies for our research. However, we established inclusion criteria, considering only studies published after 2019, that is, after the publication date of the base article. Additionally, we excluded systematic literature reviews or survey studies from our analysis. The 10 remaining studies were evaluated for their degree of similarity to the base article, represented by the similarity index Similar based-Paper (SbP), ranging from 12% to 100%. The higher the SbP value, the greater the similarity between the article and the base work (previous paper (BORGLI et al., 2020)), which is relevant to the results obtained in our research.

To gather related works for our paper, each article was thoroughly reviewed based on the following parameters: `Study` name and year, `Task` performed, CNN `Architecture` used, `Methodology` approach, `Dataset` and `SbP`. These parameters were used to assess and categorize the papers, ensuring that they align with the focus and objectives of our research. By analyzing each article based on these criteria, we were able to identify and select relevant works that contribute to our study.

Table 1 shows several studies in the context of gastrointestinal endoscopy. The studies cover a range of tasks such as polyp classification, segmentation, detection, localization,

[2] Available at: <https://www.connectedpapers.com/main>.

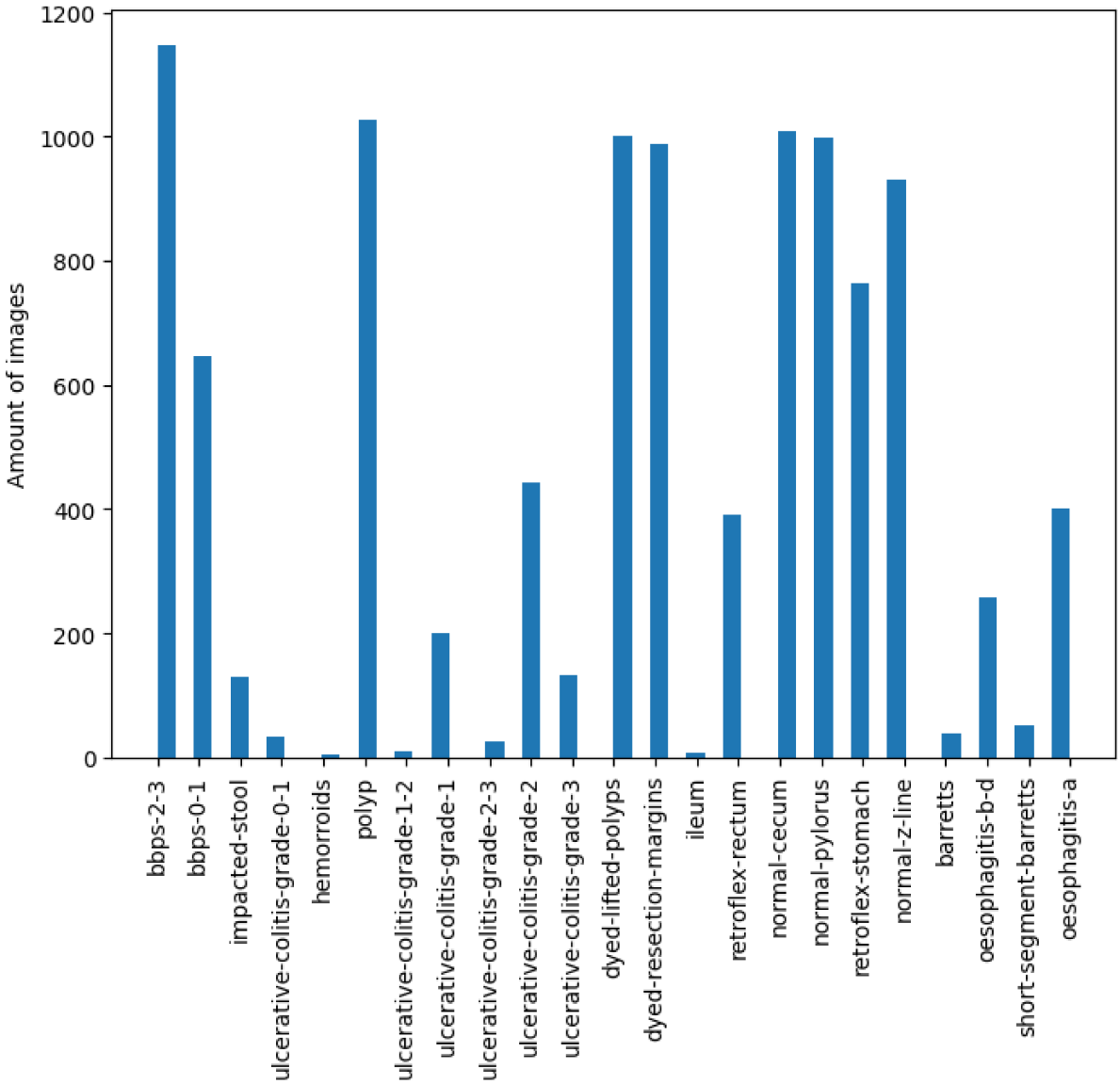


Figure 9 – Instance distribution by class.

and abnormality identification. Various deep learning architectures, including CNNs like ResNet-152, DenseNet-161, U-Net, Pix2Pix and HGANet are utilized in these studies. Different methodologies and techniques such as Fuzzy C-Means Clustering, ResUNet, Conditional Random Fields, Test-Time Augmentation, and Adversarial Training are also employed. Multiple datasets are used for evaluation, including the HyperKvasir and the EAD2019 datasets. The achieved SbP rank is used as a performance metric, with higher values indicating better results, which means is most similar to paper (BORGLI et al., 2020).

These studies provided valuable insights and advancements in leveraging deep learning and clustering techniques for gastrointestinal endoscopy. They contribute to the development of automated systems for polyp classification, segmentation, and abnormality

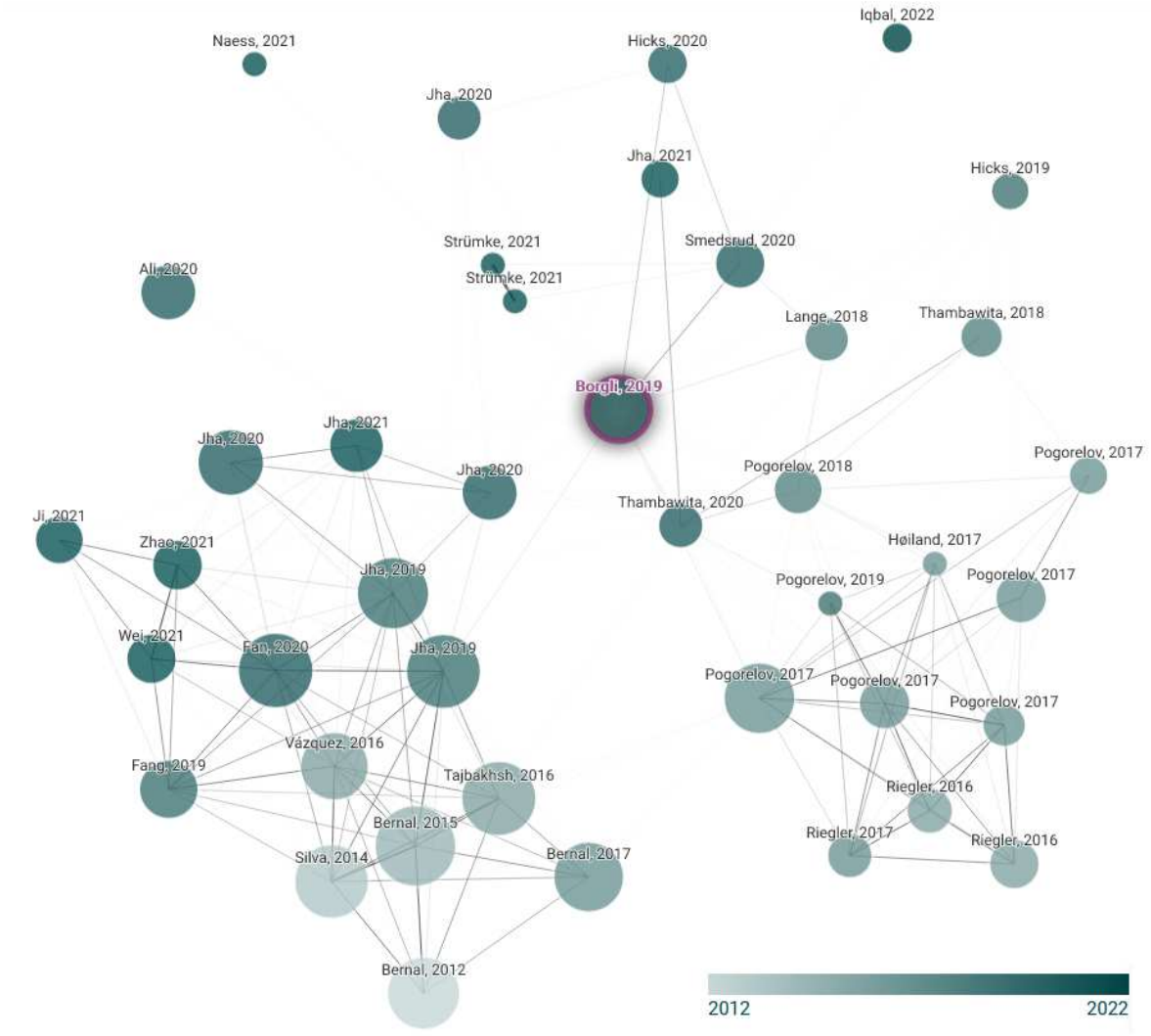


Figure 10 – Graph for connected papers to (BORGLI et al., 2020), which uses HyperKvasir image and video dataset for gastrointestinal endoscopy. In the image it shows "Borgli, 2019" because the system considered the year in which the base was created and not the year of publication which is 2020.

detection, which can improve the efficiency and accuracy of medical diagnoses. Besides that, these papers served as the foundation for our study, as they utilized various CNNs for different approaches to GI problems.

3.1.3 Background model

In (BORGLI et al., 2020), the authors proposed a classification model represented in Figure 11. The model is composed of pre-trained CNNs, DenseNet161 and ResNet152. Each CNN model is a function, M, composed of a set of subfunctions (convolution, pooling, batch normalization, softmax, optimizer, etc.) which, in this case, given an input image $\overrightarrow{\chi}$, a learning rate value η and the number of epochs e, returns an output

Table 1 – State of the art summarization considering GI dataset for medical diagnoses.

Study	**Task**	**Architecture**	**Methodology**	**Dataset**	**SbP**
HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy (BORGLI et al., 2020)	Gastrointestinal polyp Classification	Pre-trained (ResNet-50, ResNet-152, DenseNet-161 Averaged ResNet-152 + DenseNet-161, ResNet-152 + DenseNet-161+MLP	Five different deep CNN were trained and evaluated using standard classification metrics.	HyperKvasir	100
Kvasir-SEG: A Segmented Polyp Dataset (JHA et al., 2020)	Gastrointestinal polyp Segmentation	Fuzzy c-mean clustering, ResUNet CNN	Preprocessing steps, FCM algorithm, Data augmentation, ResU-Net implementation details, Qualitative comparison of FCM clustering and ResUNet results	Kvasir-SEG	19.7
The endotect 2020 challenge: evaluation and Comparison of Classification, Segmentation and Inference time for endoscopy (HICKS et al., 2021)	Gastrointestinal polyp detection and segmentation	CNN (ResNet-152), CNN (Mask Scoring R-CNN, DeepLab V3+)	Automatic segmentation of polyps, Real-time analysis	HyperKvasir	24.8
An Extensive Study on Cross-dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal tract Abnormality Classification (THAMBAWITA et al., 2020)	Gastrointestinal tract diseases Identification and Segmentation	CNN (ResNet-152, Dense Net-161) and addtional MLP	GF-based approaches, Deep neural networks, Cross-dataset evaluations, Automated identification of GI tract diseases	CVC-12k, CVC-356-plus, CVC-612-plus, 2018 Medico	20.6
Real-Time Polyp Detection, Localization and segmentation in colonoscopy Using Deep Learning (JHA et al., 2021)	Gastrointestinal polyp detection, localization and segmentation	CNN (YOLOv4 with Darknet53 backbone), Segmentation networks (Colon SegNet, UNet+ ResNet34 backbone, Deep-Labv3+, PSPNet, HRNet)	Object detection and localization using YOLOv4 with Dark-net53 backbone and Cross-Stage-Partial-Connections (CSP), Semantic segmentation + different UNet, Deep-Labv3+, PSPNet, HRNet	Kvasir-SEG	13.8
Medico Multimedia Task at Media Eval 2020: Automatic Polyp Segmentation (JHA et al., 2020)	Gastrointestinal polyp segmentation	CNN, Dice similarity coefficient (DSC) and mean Intersection over Union (mIoU)	Algorithm Speed Efficiency, Framesper-second (FPS) while segment colonoscopic images	Kvasir-Seg	12.8
An objective comparison of detection and segmentation algorithms for artefactsin clinical endoscopy (ALI et al., 2020)	Hollow-organs generalization, detection and segmentation	Mask R-CNN, RetinaNet, Cascade R-CNN, DeepLabV3	Transfer learning, ensemble techniques, out-of-sample generalization, 2-training separate batches, 7 prevalent artefact types	EAD2019 (2192 unique video, 475 video frames + mask annotations, addi= tional 195, 122, and 51 videos)	12
A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random Field and Test-Time Augmentation (JHA et al., 2021)	Gastrointestinal polyp Segmentation	CNN ResUNet++	Conditional Random Field (CRF) and TestTime Augmentation (TTA), Dice coefficient (DSC), Intersection over Union (IoU), mean IoU (mIoU), AUC-ROCand data augmentation	Kvasir-SEG, CVC-ClinicDB, CVC-ColonDB, ETIS-Larib Polyp DB, ASU-Mayo Clinic Colonoscopy Video Database, CVC-Video ClinicDB	13.1
Pyramidal segmentation of Medical Images using Adversarial Training (NAESS et al., 2021)	Gastrointestinal polyp Segmentation and localization	U-Net and Pix2Pix	Learning to segment within several grids, Grid augmentation, Cross-data training and testing	Kvasir-SEG (validation, testing), CVCClinic DB (testing)	12.7
Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images (IQBAL et al., 2022)	Gastrointestinal polyp abnormality identification	DCNN (HGANet)	HGANet with multiple routes, various image resolutions, and several convolutional layers. Pre-processing involves cropping, downsampling and removal of undesired artifacts. Augmentation techniques are applied to balance the classes.	Kvasir-Capsule	12.4

$\overrightarrow{P}$, according to the Equation 2:

$$\overrightarrow{P}_i = M(\overrightarrow{\chi}, \eta, e) \tag{2}$$

The output, $\overrightarrow{P}$, is a probability vector that indicates the probability that $\overrightarrow{v}_i$ belongs to one of the classes of the problem. The vector $\overrightarrow{P} = [c_1, c_2, c_3, \cdots, c_{23}]$, where $c \in C$ for each GI class and $|C| = 23$. Given a dataset with 10639 gastrointestinal images,

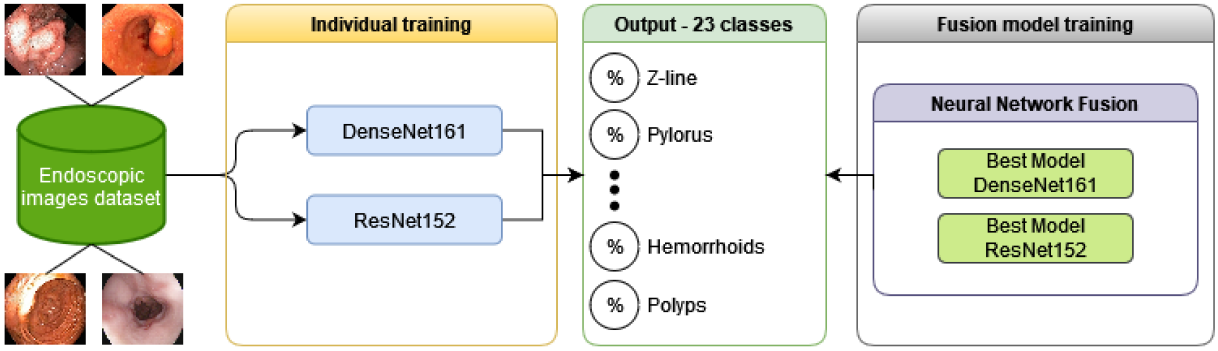


Figure 11 – Previous model (BORGLI et al., 2020) for CNN models training and fusion process.

$\overrightarrow{\chi}$, divided into two splits, one set with 5315 images and another set with 5324 images, the validation split is used in testing. The model proposed in (BORGLI et al., 2020) alternates sets for training and validation. The final response of the model is the average of the results of the two splits. For each split, the authors trained the CNNs models M_1 and M_2, respectively, DenseNet161 and ResNet152, using $\eta = 0.001$, $e = 50$, the optimizer Stochastic Gradient Descent (SGD) and batch size 32; M_1 and M_2 generated the responses $\overrightarrow{P}_1$ and $\overrightarrow{P}_2$, respectively. After training, the best weight set $\overrightarrow{w}$ for each model was found and the best-trained models M_1^b and M_2^b are saved. Using the trained models, a model M^ν is created:

$$\overrightarrow{P}_i^{\nu} = M^{\nu}(\overrightarrow{\chi}, \eta, e, M_1^b, M_2^b), \tag{3}$$

since $\overrightarrow{P}_i^{\nu} = \frac{(\overrightarrow{P}_1^b + \overrightarrow{P}_2^b)}{2}$, $\overrightarrow{P}_1^b$ and $\overrightarrow{P}_2^b$ are output of M_1^b and M_2^b, respectively.

3.2 Proposal

In this chapter, we have two proposals. The first proposal sought to overcome the literature results using fewer GPU resources than the second proposal, which is an expansion of what the authors propose in (BORGLI et al., 2020). The second proposal sought to verify whether other CNN fusions, even using more GPU resources, outperformed the literature results. The dataset for training and testing contains 10639 gastrointestinal images and has two divisions, one part with 5315 images and another part with 5324 images. The CNNs models used are accessible through the Pytorch framework.

3.2.1 Fusion and ensemble processes

Our first proposal was to individually train a set of CNNs models $\{M_1, M_2, M_3, ..., M_n\}$ and get their respective answers $\{\overrightarrow{P}_1, \overrightarrow{P}_2, \cdots, \overrightarrow{P}_n\}$, for $n = 7$. The CNNs were DenseNet121,

DenseNet161, DenseNet201, EfficientNet_b0, MobileNetV2, ResNet152 e VGG16. Pre-trained models in *ImageNet-1K*, available in *PyTorch*, were used. To minimize the problem of class imbalance, data augmentation was used. Each model was trained adopting the following values for $\eta = \{0.0001, 0.0003, 0.0005, 0.001, 0.003, 0.005\}$, $e = 50$, SGD and batch size 32, as shown in Figure 12. In the set of values of η there is the present value in (BORGLI et al., 2020) and some higher and lower values. The loss function used during training was cross entropy. After training, each model was tested and the responses were fused. Each network could participate or not in the fusion. During the tests, the networks' responses were fused, forming all possible combinations between the seven networks. Fusions occur between trained models with the same learning rate, the total of fusions were $2^n \times 6$.

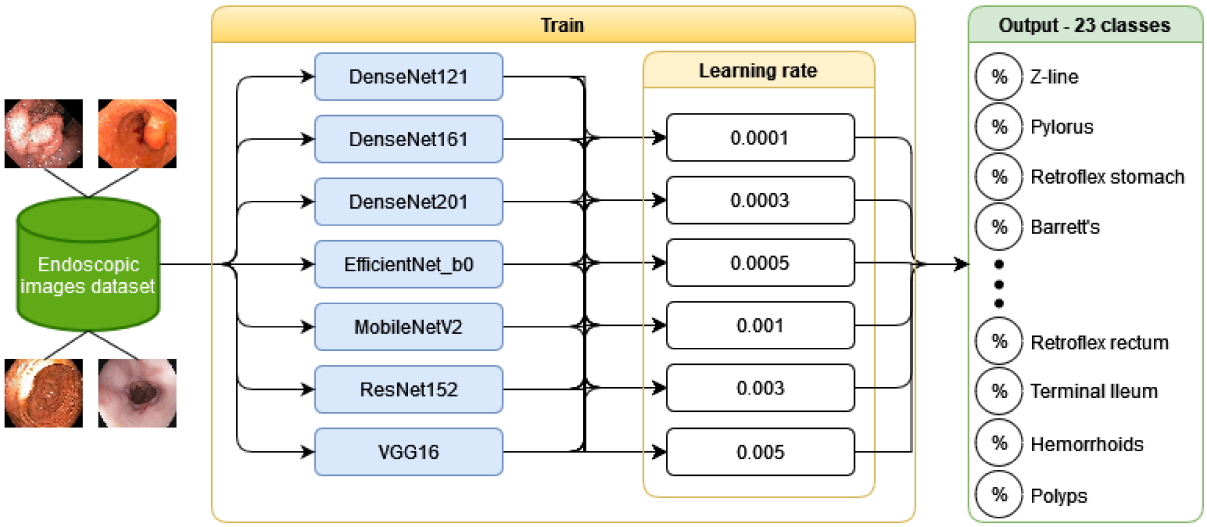
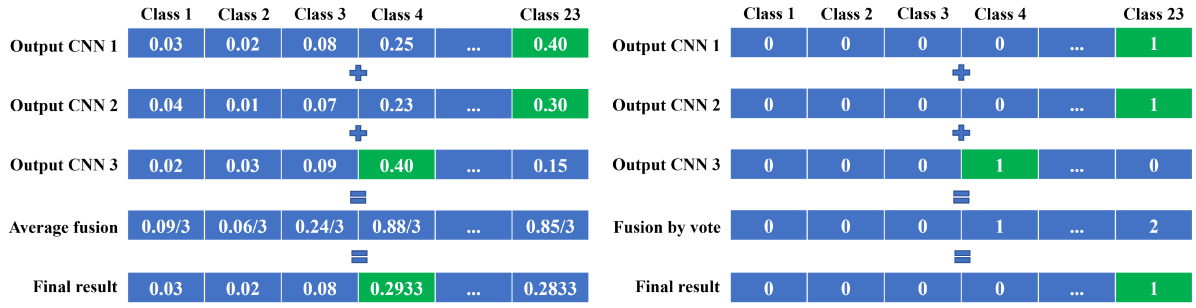


Figure 12 – Our proposal for ensemble architecture for training and fusion process.

For this proposal, two fusion alternatives were analyzed, by average and by voting. In average fusion, each trained model, M_i, generates an output $\overrightarrow{P}_i$ and the fusion is given by $\overrightarrow{P}_\tau = \frac{1}{n}\sum_{i=1}^{n} \overrightarrow{P}_i$, where n is the number of models of CNNs and $\overrightarrow{P}_\tau$ is the average output of the models that make up the fusion, according to Figure 13 (a). In fusion by voting, considering the values in Figure 13 (a), each network that makes up the fusion votes in the class that receives the highest percentage of probability, as shown in Figure 13 (b). In case of a tie, the first tiebreaker considered the number of times the class was in first and second place in the voting. If the tie remains, among the classes that met the first tiebreaker criterion, the one with the highest percentage of probability is chosen.

In our second proposal, illustrated in Figure 14, we used the model combination strategy presented by (BORGLI et al., 2020), but we performed tests with other CNN models. Models were created with different fusions of pre-trained CNNs, so M^ν was the composition of best models $\{M_1^b, M_2^b, \cdots, M_n^b\}$, for different values of $2 \leq n \leq 7$, in Figure 14 presents a fusion of models. Each chosen model was trained individually for 50 epochs.



(a) Fusion by mean scheme. (b) Fusion by voting scheme.

Figure 13 – Different fusion schemes for combining models.

The best trained were fused to form a new model, which was trained for another 50 epochs.

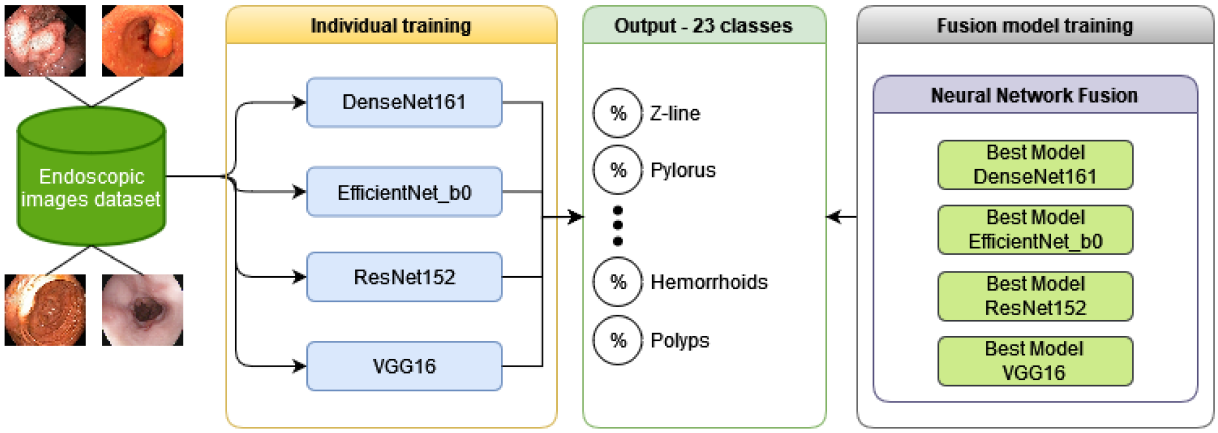


Figure 14 – Representation of the model training process. Models are trained individually. The best trained models are fused and retrained.

In the first proposal, with trained models, it is possible to perform fusions between models without having to train them again. In the second proposal, each new fusion generates a new model that needs to be trained.

3.2.2 Evaluation methodology

Our approach involved the utilization of four metrics - precision, recall, F1-score, and Matthews correlation coefficient (MCC) – to evaluate the performance of our model and gain valuable insights. Additionally, we employed both macro and micro averages to further analyze the overall performance of our model (SARKAR; BALI; SHARMA, 2018).

In classification tasks, true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are commonly used terms that represent the outcomes of the predictions made by a model. TP refers to correct predictions of the positive class, where

the model identifies positive instances correctly. TN represents correct predictions of the negative class, where the model identifies negative instances correctly. FP refers to incorrect predictions of the positive class, where the model identifies negative instances as positive. FN represents incorrect predictions of the negative class, where the model identifies positive instances as negative (SARKAR; BALI; SHARMA, 2018).

Matthews correlation coefficient, which is a measure of the quality of binary (two-class) classification models. It takes into account TP, TN, FP and FN to provide a balanced assessment of the model's performance, as shown in Equation 4. The MCC ranges from $[-1, +1]$, where a value of $(+1)$ indicates a perfect classification, (0) indicates a random classification, and (-1) indicates a completely wrong classification. MCC values closer to $(+1)$ indicate better performance of the classification model (CHICCO; JURMAN, 2020).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

Micro-average is a method of aggregating the performance metrics across all classes in a multi-class classification problem. The formulas for micro-average precision (miP) (Eq. 5), recall (miR) (Eq. 6), and F1-score ($miF1$) (Eq. 7) are as follows (TAKAHASHI et al., 2022):

$$miP = \frac{TP}{TP + FP} \tag{5}$$

$$miR = \frac{TP}{TP + FN} \tag{6}$$

$$miF1 = \frac{2 \times (miP \times miR)}{(miP + miR)} \tag{7}$$

Macro-average, on the other hand, calculates the performance metrics for each class individually and then takes the average across all classes. The formulas for macro-average precision (maP) (Eq. 8), recall (maR) (Eq. 9), and F1-score ($maF1$) (Eq. 10) are as follows (TAKAHASHI et al., 2022):

$$maP = \frac{\sum_i^n (Precision_i)}{n} \tag{8}$$

$$maR = \frac{\sum_i^n (Recall_i)}{n} \tag{9}$$

$$maF1 = \frac{\sum_i^n (F1\_score_i)}{n} \tag{10}$$

where precision_i, recall_i, and F1-score_i represent the precision, recall, and F1-score of class i, and n is the total number of classes. By using micro and macro averaging, we can gain insights into the overall performance of the classification model, considering both the individual class performance and the overall performance across all classes.

3.3 Results and Discussion

In this section, we present a comprehensive analysis of various aspects related to CNNs and their fusion configurations. Firstly, we discuss the analysis for individual CNNs. Next, we explore the evaluation of fusion configurations and CNN performance metrics. In the experiments involving average fusion and voting fusion, 768 combinations were performed, however, in the results sections we present the nine best results. In these experiments, fusion occurs during the tests and the output layer values that define which problem class the input instance belongs to are fused.

Furthermore, we investigate the fusion with optimal training CNN models. Lastly, we conduct a comprehensive analysis of resource consumption for the CNN models. By examining these four aspects, we gain a comprehensive understanding of the individual and fused CNN models, their performance metrics, optimal training configurations, and resource requirements. This knowledge enables us to make informed decisions and design more effective and resource-efficient CNN-based systems.

3.3.1 Analysis for individual convolutional neural networks

In this section, we evaluate the individual performance of the CNN models and their effectiveness in tackling the given task. This analysis provides insights into the strengths and weaknesses of each individual model. Table 2 present differents CNNs and configurations, such as Learning Rate (LR), in addition, performance metrics, including Matthews correlation coefficient (MCC), precision, recall, and F1-score. Each row represents a different CNN model, denoted by M_1 to M_7. Each model was evaluated with 6 LR's, the best results are presented in the Table 2. Analyzing the results, it can be observed that

Table 2 – Individual convolutional neural networks results.

ID	CNN Models	LR	**Macro Average**			**Micro Average**			MCC
			Precision	Recall	F1-Score	Precision	Recall	F1-Score	
M_1	DenseNet121	0.0030	0.6149	0.6004	0.5986	0.8929	0.8929	0.8929	0.8839
M_2	DenseNet161	0.0030	0.6190	0.6016	0.6045	**0.9025**	**0.9025**	**0.9025**	**0.8942**
M_3	DenseNet201	0.0030	0.6199	0.5963	0.5991	0.8972	0.8972	0.8972	0.8884
M_4	EfficientNet_b0	0.0050	0.5952	0.6078	0.5955	0.8902	0.8902	0.8902	0.8810
M_5	MobileNetV2	0.0030	0.5999	0.5976	0.5928	0.8856	0.8856	0.8856	0.8760
M_6	ResNet152	0.0050	**0.6252**	**0.6068**	**0.6094**	0.9007	0.9007	0.9007	0.8923
M_7	VGG16	0.0030	0.5893	0.5936	0.5874	0.8846	0.8846	0.8846	0.8749

different CNN models achieve varying levels of performance across the evaluated metrics. Among the models, M_2 (DenseNet161) stands out with the highest precision, recall, F1-score, and MCC values. On the other hand, M_5 (MobileNetV2) and M_7 (VGG16) exhibit slightly lower performance in terms of precision, recall, F1-score, and MCC.

3.3.2 Analysis of fusion configurations and CNN performance metrics

In this section, we explore different fusion techniques and assess their impact on the overall performance of the system. The table 3 presents the nine best results obtained from the fusion of multiple convolutional neural networks (CNNs), ensembles, using the average method. Each row in the table represents a different fusion scenario, denoted by F_i, where multiple CNN models (M_1 to M_7) are combined, where (`Y`) indicates the presence of a model in the fusion configuration, while (`N`) denotes the absence of that model. The table also includes the LR used for each fusion scenario. The evaluation metrics used to assess the performance of the fusion approach are precision, recall, F1-score, and MCC. Analyzing the results, it can be seen that different ensemble architectures and their fusion responses generate varying levels of performance in the evaluated metrics. Among

Table 3 – Results obtained with fusion of CNN by average.

F_i	M_1	M_2	M_3	M_4	M_5	M_6	M_7	LR	**Macro Average**			**Micro Average**			MCC
									Precision	Recall	F1-Score	Precision	Recall	F1-Score	
1	N	Y	N	Y	N	Y	Y	0.0030	0.6312	0.6085	0.6084	**0.9101**	**0.9101**	**0.9101**	**0.9025**
2	Y	Y	Y	Y	N	Y	Y	0.0030	**0.6402**	0.6121	0.6143	0.9101	0.9101	0.9101	0.9025
3	N	Y	N	Y	N	Y	Y	0.0050	0.6395	**0.6180**	**0.6187**	0.9100	0.9100	0.9100	0.9023
4	Y	Y	Y	Y	Y	Y	Y	0.0030	0.6334	0.6138	0.6150	0.9100	0.9100	0.9100	0.9023
5	Y	Y	N	Y	N	Y	Y	0.0030	0.6360	0.6114	0.6118	0.9100	0.9100	0.9100	0.9023
6	Y	Y	Y	N	N	Y	Y	0.0030	0.6399	0.6105	0.6139	0.9097	0.9097	0.9097	0.9019
7	Y	Y	N	Y	N	Y	Y	0.0050	0.6265	0.6165	0.6154	0.9097	0.9097	0.9097	0.9020
8	Y	Y	Y	N	Y	Y	Y	0.0030	0.6352	0.6105	0.6123	0.9096	0.9096	0.9096	0.9018
9	N	Y	N	N	N	Y	N	0.0030	0.6347	0.6061	0.6102	0.9074	0.9074	0.9074	0.8995

the merger scenarios, F_3 displays the highest F1-score value for macro-average. This scenario combines specific CNN models (M_2, M_4, M_6 and M_7) and achieves remarkable performance in correctly classifying positive and negative instances.

These scenarios show the effectiveness of ensemble methods to improve classification accuracy. Fusion F_1 stands out, achieving relatively high values of accuracy, recall, F1-score and MCC, for micro average. This suggests that the combination of models M_2, M_4, M_6 and M_7 with an LR of $\eta = 0.003$ leads to successful predictions with high accuracy and completeness. Considering the CNN models that appeared more frequently in the fusion experiments, the models M_2, M_6 and M_7 were used in a greater number of experiments. This suggests that these models have a greater impact on the overall performance of the ensemble architectures. Table 4 shows the top nine CNN fusion using a voting mechanism where each model in the ensemble makes an independent prediction, and the final prediction is based on the highest number of votes. Upon analyzing the results, it is evident that the performance of the fusion models varies depending on the specific combination of CNN models used. F_2 stands out as it achieves the highest values in terms of F1-score for macro (0.6158), F1-score for micro average (0.9089) and MCC (0.9012). This combination includes models M_1, M_2, M_3, M_4, and M_7, indicating that these models contribute significantly to the overall performance of the fusion model.

Table 4 – Results obtained with CNN fusion by vote.

F_i	M_1	M_2	M_3	M_4	M_5	M_6	M_7	LR	**Macro Average**			**Micro Average**			MCC
									Precision	Recall	F1-Score	Precision	Recall	F1 Score	
1	N	Y	Y	Y	N	Y	Y	0.0030	0.6359	0.6093	0.6102	0.9089	0.9089	0.9089	0.9011
2	Y	Y	Y	Y	N	N	Y	0.0030	**0.6416**	**0.6138**	**0.6158**	**0.9089**	**0.9089**	**0.9089**	**0.9012**
3	Y	Y	N	Y	N	Y	Y	0.0030	0.6308	0.6125	0.6125	0.9088	0.9088	0.9088	0.9011
4	Y	Y	Y	N	N	Y	Y	0.0030	0.6403	0.6124	0.6157	0.9088	0.9088	0.9088	0.9010
5	N	Y	Y	Y	Y	Y	Y	0.0030	0.6264	0.6104	0.6102	0.9084	0.9084	0.9084	0.9006
6	Y	Y	Y	Y	N	Y	N	0.0030	0.6288	0.6079	0.6080	0.9084	0.9084	0.9084	0.9005
7	Y	Y	Y	Y	Y	Y	Y	0.0030	0.6290	0.6096	0.6097	0.9082	0.9082	0.9082	0.9003
8	N	Y	N	N	N	Y	Y	0.0030	0.6376	0.6112	0.6146	0.9081	0.9081	0.9081	0.9002
9	N	Y	N	N	N	Y	N	0.0050	0.6308	0.6121	0.6147	0.9038	0.9038	0.9038	0.8956

3.3.3 Fusion with optimal training CNN models

In this section, we explore the integration of specific CNN models to further enhance the system's performance and achieve superior results. The results refer to proposal two, presented in Section 3.2.1. Table 5 presents the results of the fusion of trained CNN models using different combinations. In our experiments, in F_4^b, considering the amount of CNNs that make up the fusion, we had the best performance compared to other combinations, considering both micro-average, F1-score (0.9126) and MCC (0.9051). Overall, the analysis of the fusion results indicates that the combinations involving M_2^b,

Table 5 – Fusion with trained CNN models.* Refers to the combination of CNNs presented in (BORGLI et al., 2020).

F_i^b	M_1^b	M_2^b	M_3^b	M_4^b	M_5^b	M_6^b	M_7^b	LR	**Macro Average**			**Micro Average**			MCC
									Precision	Recall	F1-Score	Precision	Recall	F1-Score	
1*	N	Y	N	N	N	Y	N	0.0010	0.6330	0.6150	0.6170	0.9100	0.9100	0.9100	0.9020
2	N	Y	N	N	N	N	Y	0.0030	**0.6340**	0.6172	0.6202	0.9081	0.9081	0.9081	0.9002
3	N	Y	N	N	N	Y	Y	0.0030	0.6339	**0.6212**	**0.6246**	0.9121	0.9121	0.9121	0.9046
4	N	Y	N	Y	N	Y	Y	0.0030	0.6328	0.6211	0.6232	**0.9126**	**0.9126**	**0.9126**	**0.9051**
5	N	Y	Y	Y	N	Y	Y	0.0030	0.6298	0.6215	0.6227	0.9124	0.9124	0.9124	0.9049
6	Y	Y	Y	Y	N	N	Y	0.0030	0.6273	0.6163	0.6178	0.9110	0.9110	0.9110	0.9034
7	Y	Y	Y	Y	N	Y	Y	0.0030	0.6290	0.6193	0.6214	0.9128	0.9128	0.9128	0.9053

M_4^b, M_6^b, and M_7^b generally lead to better performance, with higher F1-scores and MCC values. The presence of M_1^b and M_5^b does not contribute significantly to the overall performance improvement.

3.3.4 Resource Consumption Analysis for CNN Models

In Table 6, we present details such as the CNN model name, GPU model used for execution, GPU RAM capacity, execution time in minutes, and the number of parameters for each model. To measure the execution time and GPU consumption, the `timeit` module and the `psutil` library were used, respectively. All network models, M_1 to M_7, utilize the Tesla V100-SXM2-16GB GPU model. Additionally, the F_1^{b*} model (BORGLI et al., 2020) employs the Tesla V100-SMX2-16GB, while fusion models F_2^b to F_7^b, for Table 5 and Table 6, utilize the Nvidia A100-SXM-40GB GPU model. Models F_2^b to F_7^b consume more GPU RAM and more execution time, so they needed to be executed on another

Table 6 – Resource consumption for network models.

Individual Models					Fusion Models				
CNN models	**GPU model**	**RAM GPU (GB)**	**Execution Time (m)**	**Parameters**	**CNN models**	**GPU model**	**RAM GPU (GB)**	**Execution Time (m)**	**Parameters**
M_1	tesla v100-sxm2-16gb	6.2	92.0	6977431	F_1^{b*}	tesla v100-sxm2-16gb	15.8	113.2	84713742
M_2	tesla v100-sxm2-16gb	10.0	94.9	26522807	F_2^b	nvidia a100-sxm-40gb	15.7	55.5	160877582
M_3	tesla v100-sxm2-16gb	8.9	89.9	18137111	F_3^b	nvidia a100-sxm-40gb	22.7	82.9	219068517
M_4	tesla v100-sxm2-16gb	4.6	80.9	4037011	F_4^b	nvidia a100-sxm-40gb	24.7	98.5	223105528
M_5	tesla v100-sxm2-16gb	4.2	86.2	2253335	F_5^b	nvidia a100-sxm-40gb	31.0	89.6	241242639
M_6	tesla v100-sxm2-16gb	8.1	101.3	58190935	F_6^b	nvidia a100-sxm-40gb	28.9	110.3	190029135
M_7	tesla v100-sxm2-16gb	7.0	109.8	134354775	F_7^b	nvidia a100-sxm-40gb	35.9	124.5	248220070

GPU model.

These data allow us to analyze the computational cost associated with achieving the results presented in Table 3, Table 4, and Table 5.

In Table 3, the F_1 result, including M_2, M_4, M_6, and M_7, achieved the highest F1-score (0.9101) with the least number of models used for micro-average. The models were executed individually on the GPU, resulting in a total GPU consumption equal to the highest consumption among the individual models, which is 10GB for M_2. Thus, the proposed ensemble F_1 with CNN model averaging has a maximum GPU consumption of 10GB.

In Table 4, using the technique of fusion by vote, the ensemble F_2 achieved the highest F1-score of 0.9089 for micro and F1-score of 0.6158 for macro. The set F_2 consisted of models M_1, M_2, M_3, M_4 and M_7. The GPU consumption for the set F_1 corresponds to that of the model M_2, which is 10GB.

In Fusion with optimal training CNN models, as shown in Table 5, for F_1^{b*}, the F1-score is 0.910, which matches our proposal in Table 3. The approach in (BORGLI et al., 2020), F_1^{b*}, requires 15.8GB of GPU, as the best models M_2^b and M_6^b are trained together. Building upon the combination of models proposed in (BORGLI et al., 2020), we introduce ensembles F_2^b to F_7^b, with F_4^b achieving the best result. Figure 15 depicts a bubble chart illustrating that we have achieved comparable results (indicated by the blue and red bubbles) when compared to the fusion model F_1^{b*} in Table 5 (green bubble), as presented in (BORGLI et al., 2020).

Our proposal F_1 in Table 3 attained the same results while utilizing 10GB of GPU, which is 36.7% less than the consumption of (BORGLI et al., 2020) with 15.8GB GPU. The purple bubbles demonstrate that our ensemble architectures using Fusion with optimal training CNN models obtain better results than (BORGLI et al., 2020), albeit at a higher GPU cost.

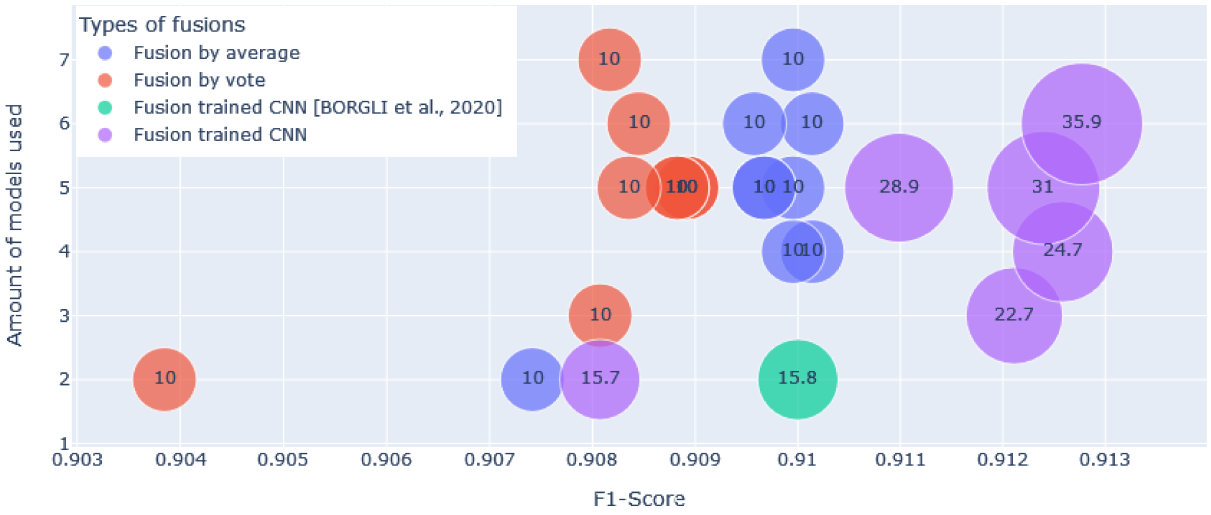


Figure 15 – Bubble chart for resource consumption analysis and comparative results. The diameter of the bubbles is associated with GPU consumption.

Chapter 4

Fusion of CNNs for Medical Images Landmarks Detection

With technological advances, AI tools, such as CNN models, have been used in the task of detecting landmarks. These models automate landmark detection, reducing the need for manual intervention, saving time and resources. CNNs can be applied in different contexts and types of images, from cephalograms to cancer detection and monitoring (CANDEMIR et al., 2013; JAEGER et al., 2014; GIRSHICK et al., 2015; LITJENS et al., 2017; RUNDO et al., 2020; BAYOUDH et al., 2021).

The detection of landmarks in lateral cephalograms is essential for orthodontists, as it facilitates the diagnosis and monitoring of the evolution of craniofacial conditions, in addition, it facilitates treatment planning such as surgeries and implants (ZENG et al., 2021). The detection of landmarks is essential for the analysis of lung regions, being crucial in the diagnosis of respiratory diseases, such as COVID-19 (GIEŁCZYK et al., 2022; GAGGION et al., 2022).

In this chapter we focus on performing CNN fusion to detect landmarks in medical images. In works such as Borgli et al. (BORGLI et al., 2020) and Bayoudh et al. (BAYOUDH et al., 2021), the authors discuss the improvement of CNNs in solving problems through strategies such as CNN committees and CNN fusions. We use datasets containing lateral cephalograms, hand x-rays and lung x-rays, with the aim of detecting landmarks. Furthermore, in line with green computing, this work presents the energy consumption and carbon emissions of the studied CNN models.

4.1 Background

With improvements in GPU technology, several convolutional neural network models have been proposed in the last decade. In this work we explore the use of the models proposed by (CHEN et al., 2019) and (ZHU et al., 2021; ZHU et al., 2022). The (CHEN et al., 2019) model was used to detect points in lateral cephalograms. The (ZHU et al.,

2021; ZHU et al., 2022) model, described by the authors as a universal model, was used to detect points on lateral cephalograms, hand radiographs, and lung radiographic images. We propose to fuse both models with the aim of obtaining a universal model that achieves better results in landmark detection.

4.1.1 Datasets for landmark detection

The dataset used contains 400 images of lateral cephalograms[1]. This dataset was used in *IEEE 2015 ISBI Grand Challenge#1*. In this dataset, the goal is to detect 19 landmarks, as seen in Figure 16 (LINDNER et al., 2016). A dataset containing hand

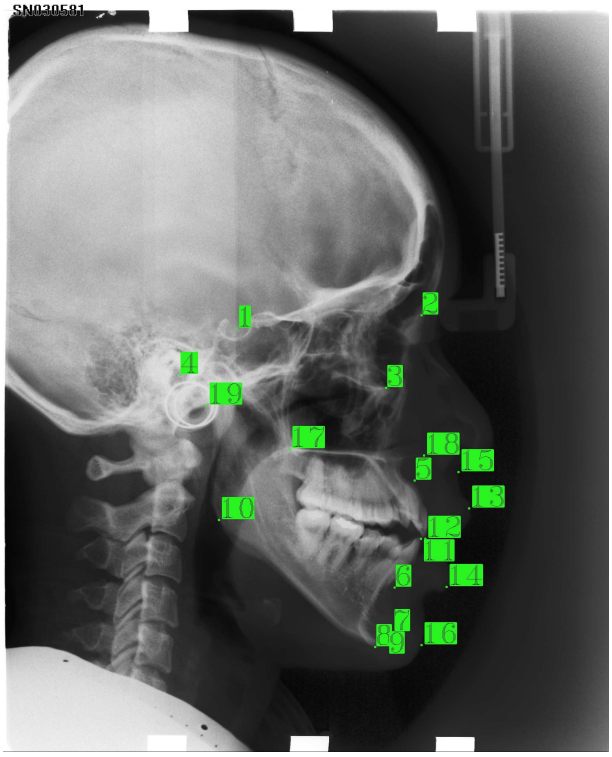


Figure 16 – Lateral cephalogram with 19 landmarks.

x-rays[2] was also used. It contains 1390 images. In this dataset, the goal is is to detect 37 reference points, as shown in Figure 17. The third dataset used in this work contains 500 x-rays images of the lung[3]. The images were labeled by radiologists (JAEGER et al., 2013; CANDEMIR et al., 2013; JAEGER et al., 2014). In these images, six landmarks are relevant, as shown Figure 18.

[1] <https://figshare.com/s/37ec464af8e81ae6ebbf>
[2] <https://ipilab.usc.edu/research/baaweb/>
[3] <https://www.kaggle.com/datasets/kmader/pulmonary-chest-xray-abnormalities>

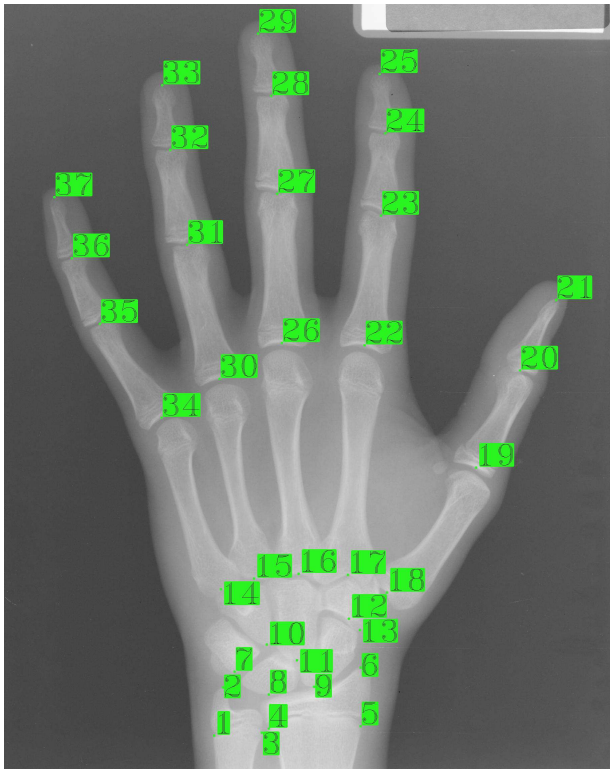


Figure 17 – Example of hand radiography with the 37 landmarks.

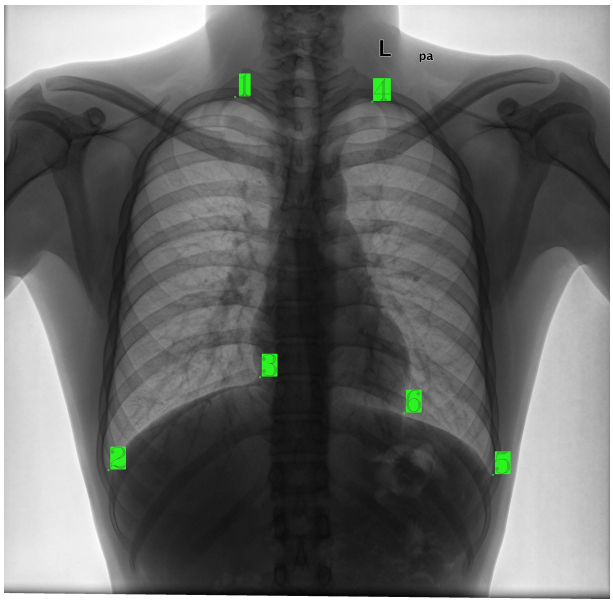


Figure 18 – Example of lung X-ray image with the 6 landmarks.

4.2 Deep learning models for object and region detection

The human brain is adapted to identify and interpret visual patterns, allowing us to recognize objects efficiently and accurately. The human ability to detect objects in images

is crucial for visual perception and understanding, enabling us to interpret environments and make appropriate decisions based on available visual information. In areas such as security and medicine, humans (experts) need to detect objects, people, organs or regions of interest to make decisions (GONZALEZ; WOODS, 2009; REDMON et al., 2016).

Computational advancement allows many object detection tasks to be performed by computational algorithms such as deep neural networks. With the use of deep neural networks, it becomes possible to extract complex features and learn hierarchical representations of images. The ability of CNNs to automatically learn objects' discriminating features, combined with training on large labeled datasets, significantly boosts the ability to detect objects in images. Detectors based on CNNs can identify objects of different sizes and poses and in different contexts (REDMON et al., 2016; LITJENS et al., 2017; BOCHKOVSKIY; WANG; LIAO, 2020; PAYER et al., 2019).

You Only Look Once (YOLO) is a model of deep neural networks for objects identification of different sizes and positions within the image, without the need to apply a sliding window. The YOLO system models detection as a regression problem. It divides the image into a grid and for each grid cell predicts a number of bounding boxes, a confidence value for those boxes, and probabilities that the bounding boxes belong to a class (REDMON et al., 2016; REDMON; FARHADI, 2017; BOCHKOVSKIY; WANG; LIAO, 2020).

The YOLO models represent an important advance in the area. These models are known for their efficiency and accuracy in real-time detection, allowing fast and accurate identification of objects in images and videos. These models have positively impacted areas such as industrial automation and computer vision (REDMON et al., 2016; BOCHKOVSKIY; WANG; LIAO, 2020).

The various versions of YOLO are being tested on large datasets for the detection of objects such as COCO and PASCAL VOC. In these baseline works, there is no in-depth study of the use of YOLO in medical imaging. One of the problems with using YOLO in medical imaging is that it focuses more on speed for real-time detection than on accuracy. In computer analysis of medical images, the main goal is accuracy (REDMON et al., 2016; REDMON; FARHADI, 2017; REDMON; FARHADI, 2018; BOCHKOVSKIY; WANG; LIAO, 2020; WANG; BOCHKOVSKIY; LIAO, 2021; WANG; BOCHKOVSKIY; LIAO, 2023; WANG; LIAO; YEH, 2022).

Other studies such as (PAYER et al., 2019; CHEN et al., 2019; ZENG et al., 2021; ZHU et al., 2021) present techniques to detect regions in medical images. These studies are focused on accuracy.

Payer et al., in (PAYER et al., 2016) and (PAYER et al., 2019), propose the use of heat maps as a way to filter out relevant landmarks. According to the authors, the proposal incorporates the spatial configuration of anatomical landmarks in a CNN-based heat map regression framework. Thus, the studies perform well in locating landmarks and do not

rely on large datasets. In the studies, the authors propose the automatic detection of regions in hand radiography images, cephalometric images, and computed tomography of the spine.

In (CHEN et al., 2019), the authors adopt the combination of heat maps, displacement maps, and pixel regression voting to propose a deep learning model that automatically detects cephalometric landmarks with high accuracy. In (LEE et al., 2020), the authors also automatically detect cephalometric landmarks using trust regions and Bayesian convolutional neural networks. In the study proposed by (ZENG et al., 2021), the authors combine three models of CNNs to automatically detect cephalometric landmarks. In (ZHU et al., 2021) and (ZHU et al., 2022), propose a universal deep learning model for landmarks detection in medical images, including cephalometric images. In Figure 19, in the first column, it is possible to see the landmarks made by specialists. In the second column, the landmarks predicted by a CNN model. In the third column, the comparison between predicted and original landmarks. In the last column, the more yellow the points are, the greater the correspondence between the points, i.e., the distance in *mm* between the points is smaller.

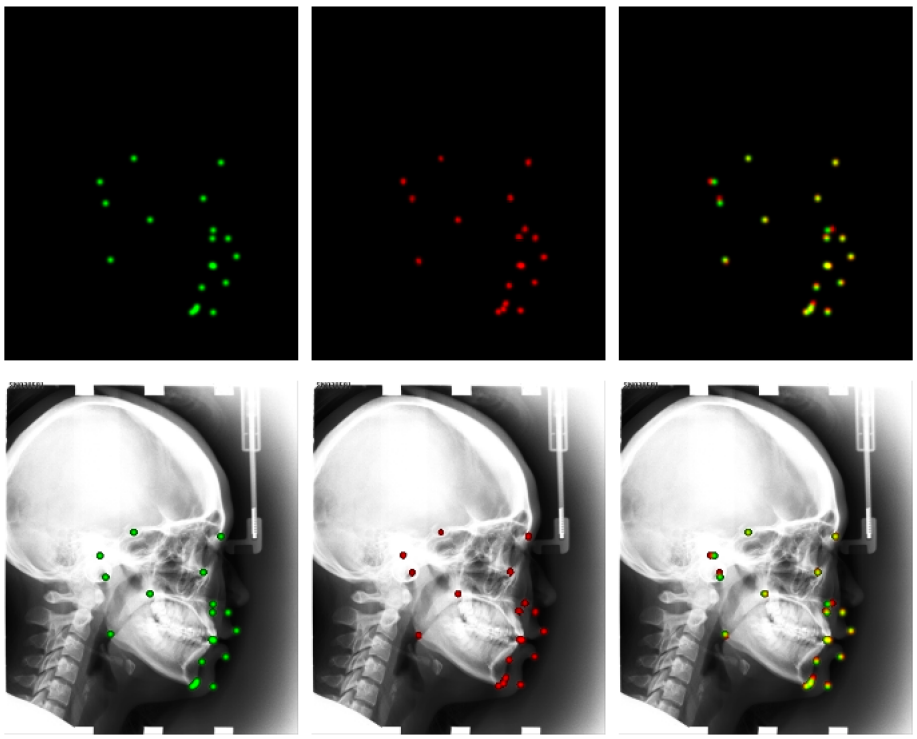


Figure 19 – Example of using CNN to detect landmarks in cephalometric images.

In a systematic review and meta-analysis presented in (MESQUITA et al., 2023), the authors report that AI techniques are adopted for the detection of cephalometric landmarks and reach an accuracy of 79% and 90%, considering an error of 2 and 3mm, respectively (MESQUITA et al., 2023). Using CNNs, there is still possibility of improve-

ment, considering the 2mm error margin. In (ZENG et al., 2021) the author mentions that 2mm is the most clinically accepted accuracy range.

A point of intersection between some studies addressed in this section is the automatic identification of landmarks in cephalometric images. According to Table 7, four studies present test results with datasets of cephalometric images. For the 2mm margin of error, the values range from 73.33 to 82.11.

Table 7 – Results presented in related works.

Studies	Cephalometric				
	MRE	SDR(%)			
	(mm)	2mm	2.5mm	3mm	4mm
Payer et al. (PAYER et al., 2019)	-	73.33	78.76	83.24	89.75
Chen et al. (CHEN et al., 2019)	1.33	80.86	87.76	92.04	96.79
Lee et al. (LEE et al., 2020)	1.53	82.11	88.63	92.28	95.96
Shu et al. (ZHU et al., 2021)	1.54	77.79	84.65	89.41	94.93

Analyzing Table 7, it was possible to integrate two studies, the proposal by Chen et al. (CHEN et al., 2019) with the proposal by Shu et al. (ZHU et al., 2021), to generate a new universal model for detecting landmarks in medical images. In this Chapter, in the next sections, we will explore the studies presented by (CHEN et al., 2019) and (ZHU et al., 2021).

4.2.1 Model Attentive Feature Pyramid Fusion and Regression-Voting

In (CHEN et al., 2019) the author proposes a framework[4] composed of the modules Feature Extraction, Attentive Feature Pyramid Fusion (AFPF) and Prediction, as illustrated in Figure 20.

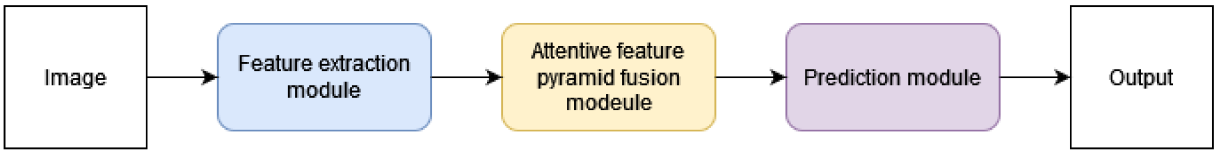


Figure 20 – Model attentive feature pyramid Fusion and regression-Voting processing steps.

The feature extraction module receives input images with dimensions $(h \times w)$, in RGB format. The images are passed through layers extracted from VGG19, available in the Pytorch framework. The layers can be divided into 4 blocks, as show in Figure 21, each block contains convolutions and max pooling. The output data of one block becomes

[4] <https://github.com/runnanchen/Anatomic-Landmark-Detection>

the input data of the next block. As the image passes through the blocks, more features are generated and the dimension of each feature is reduced. The outputs produced by each block are adjusted with upsampling and 1×1 kernel convolutions, generating 64 features of dimensions $(h/4, w/4)$. The outputs of the blocks, with the same dimensions and the same number of features, are concatenated and generate a feature map, which is the output value of the feature extraction module (CHEN et al., 2019).

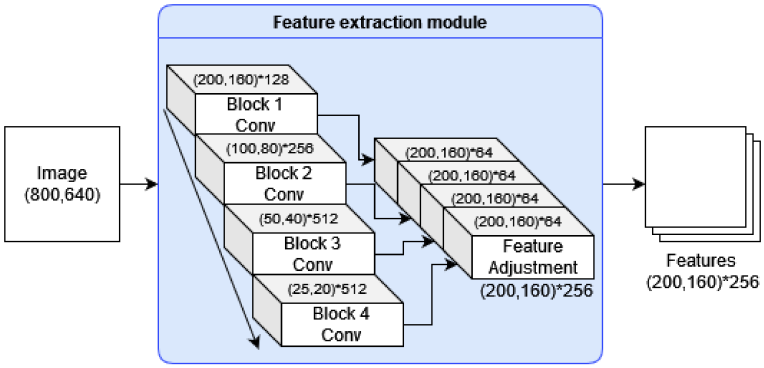


Figure 21 – Processing steps of feature extraction module. The input image is subjected to the first of four convolution blocks. The dimensions of the outputs of each convolution block are adjusted and concatenated.

Feature maps are processed in the AFPF module. In this module, feature maps undergo dilated convolutions to create the feature pyramid. This convolution inserts space between each kernel value, it enlarges the kernel. Dilated convolutions allow them to cover a larger region of the input and, therefore, improve the accuracy of the reference point estimation. Additionally, AFPF contains a mechanism that learns attention weights for each reference point (CHEN et al., 2019).

The AFPF module produces a tensor T with the size $(3n, h, w)$, $3n$ is composed of n heat maps and $2n$ offset maps; n is the number of landmarks. Heat maps H are used to delimit the approximate area of the reference point, H is $H(x, y)$. Offset maps O are regressors to locate the precise position of the reference point, O is $O(o_1, o_2)$ with $o_1 = O(x, y)$ and $o_2 = O(x, y)$ (CHEN et al., 2019; PAPANDREOU et al., 2017). In the prediction module, heat maps and offset maps are combined to predict the locations of reference points (CHEN et al., 2019; PAPANDREOU et al., 2017).

During model training, for each pixel location p_i, such that $p_i = (x, y)$, and landmarks l_n, being $l_n = (x, y)$, the probability of p_i being within a circular domain of radius R is calculated such that $H_n(p_i)$ is 1 if $p_i - l_n \leq R$ and 0 otherwise. The loss function L_h is defined as the average logistic losses between the predicted heat map and the actual heat map. Displacement maps are used to predict the $2D$ displacement vector between p_i and l_n, such that $O_n(p_i) = (l_n - p_i)/R$. The L_o loss function is defined as the $L1$ loss between the predicted and actual offset maps. The loss is calculated only for points within radius R and not across all pixels of the displacement maps. The final loss function is as

follows (CHEN et al., 2019; PAPANDREOU et al., 2017):

$$L(\theta) = \alpha L_h(\theta) + (1 - \alpha)L_o(\theta), \tag{11}$$

The value of θ corresponds to the values of offset maps and heat maps.

4.2.2 Model Universal Anatomical Landmark Detection

The model proposed by authors (ZHU et al., 2021; ZHU et al., 2022) is called the universal model[5] for detecting anatomical landmarks. According to the authors, datasets from different types of medical images can share related characteristics, such as anatomical landmarks, show in Figure 22. The authors cite features such as likely location on corners,

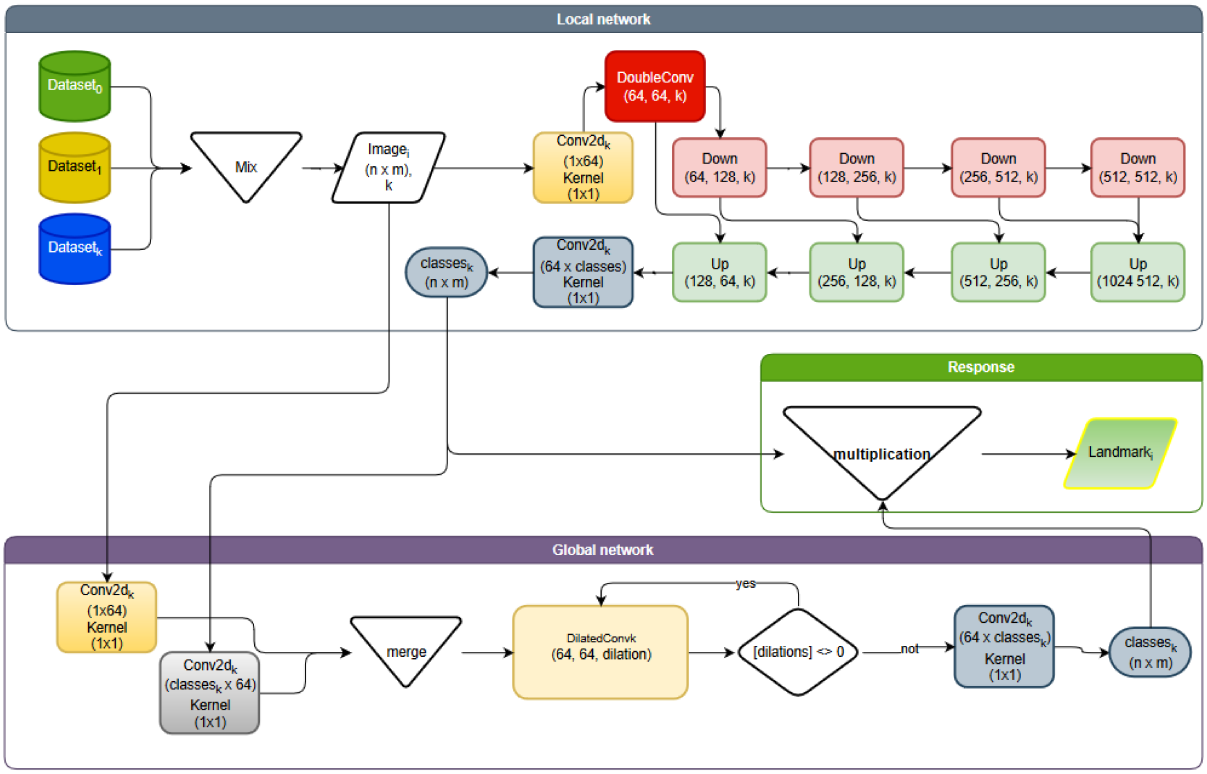


Figure 22 – Details of the universal anatomical landmark detection model. Adapted from: (ZHU et al., 2021)

edges, ends of curves, surfaces, etc. The model seeks to leverage common knowledge across datasets to achieve greater efficiency (ZHU et al., 2021; ZHU et al., 2022). The model automatically detects points of interest in medical images. The model's final response results from the fusion of the response of a local module L and a global module G, as shown in Figure 23. The local module is based on U-Net (ZHU et al., 2021; ZHU et al., 2022).

[5] <https://github.com/MIRACLE-Center/YOLO_Universal_Anatomical_Landmark_Detection>

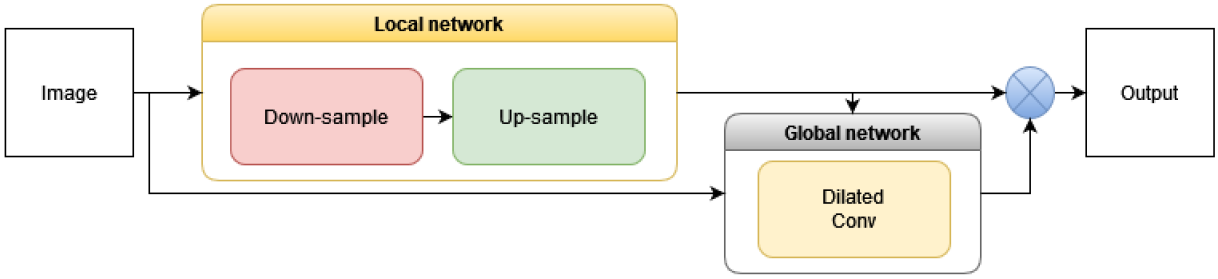


Figure 23 – Model universal anatomical landmark detection processing steps. The input image is submitted to the local network and the global network. The local network output is processed by global network. The local and global network outputs are combined to generate the final answer.

The local module comprises a down-sample and an up-sample, as illustrated in Figure 24. The down-sample blocks are composed of convolutions and maximum pooling, as the input image is down-sampled, features with smaller dimensions are created in greater quantities. The up block receives an input from the previous up block and an input from the down block. As data passes through the up blocks, the number of features decreases and the dimensions of the features increase. The relationship that exists between the down and up blocks is a connection to reestablish the dimensions of the data at each step, until reaching the original dimensions, which are the dimensions of the input (ZHU et al., 2021; ZHU et al., 2022).

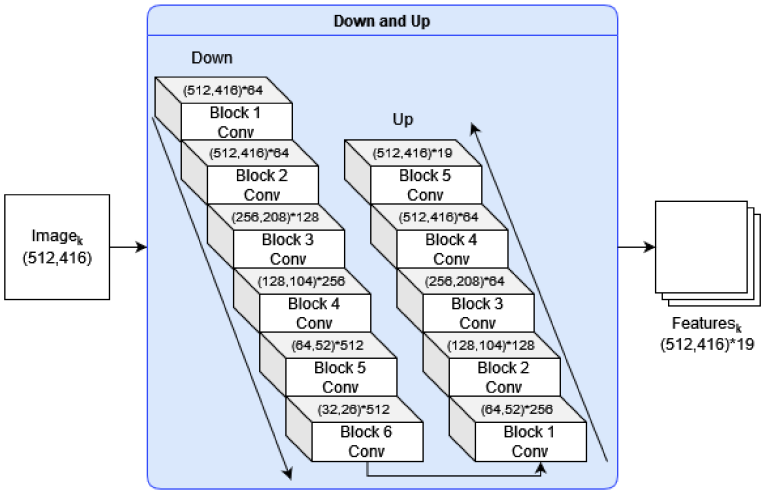


Figure 24 – Down and up convolution blocks of the universal anatomical landmark detection model.

Before starting to train the model proposed by (ZHU et al., 2021; ZHU et al., 2022), the landmarks are transformed into heat maps, show in Figure 25, using a Gaussian function (Equation 12). Given a dataset D_j that contains images S_k, $S_k = S_k(x, y)$, and each image has corresponding landmarks l_n, $l_n = l_n(x, y)$, such that j is number of

datasets, k is the number of images in a dataset and n is the number of landmarks in an image, the Gaussian function is as follows (ZHU et al., 2021; ZHU et al., 2022):

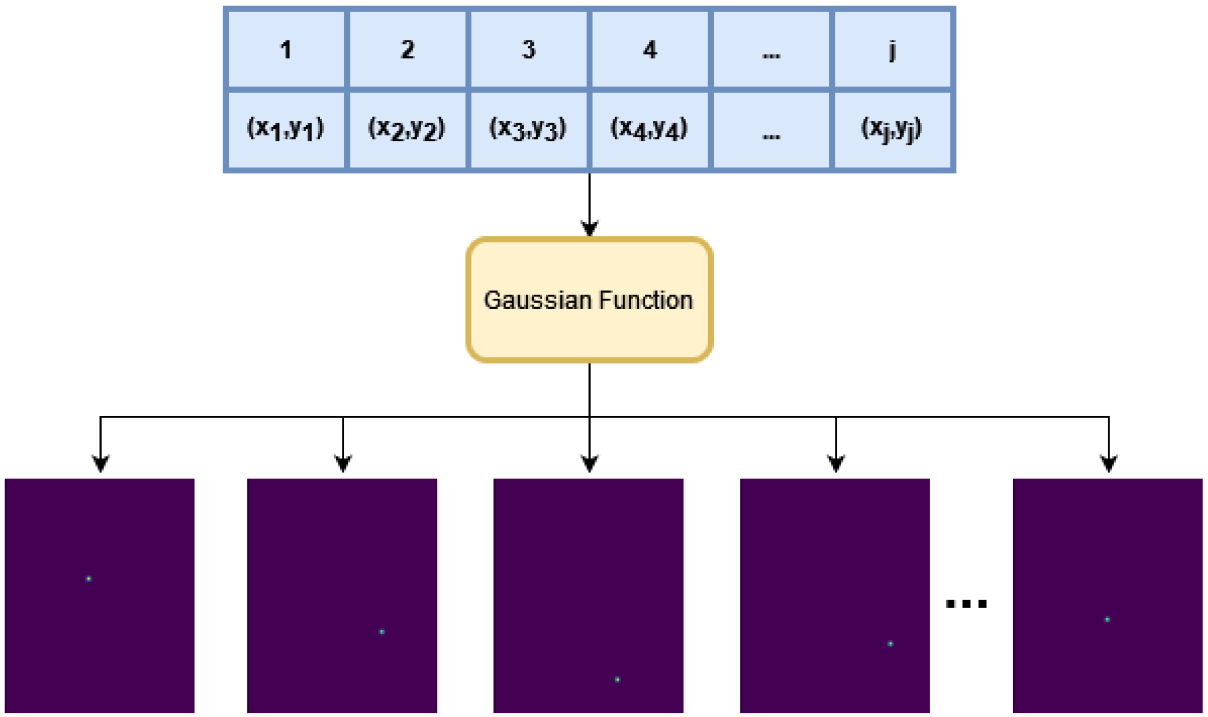


Figure 25 – Transformation of landmarks made by experts into heat maps.

$$Y_{kn} = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x_k - x_{kn})^2 + (y_k - y_{kn})^2}{2\sigma^2}, \tag{12}$$

where Y_{kn} is the n heat maps for each landmarks contained in k images.

During training, models G and L produce, respectively, tensors T^g and T^l of size (n, h, w). These tensors are combined resulting in a final tensor F – therefore $F_k = T_k^g \odot T_k^l$. The model loss is calculated by comparing the tensor F with the heat maps Y_{kn}, show in Equation 13 (ZHU et al., 2021; ZHU et al., 2022).

$$L_k = \sum_{y \in Y_k, f \in F_k} -y \log f - (1 - y) \log(1 - f) \tag{13}$$

After training, during testing, for each image in the dataset the model generates a set of landmarks, F_k, heat maps, from which the positions of the highest values are extracted, the landmarks, depicted in Figure 26.

4.3 Proposals

In this chapter, seeking to improve the performance of models for landmark detection, two proposals are reported. In the first proposal, the two complete models are fused,

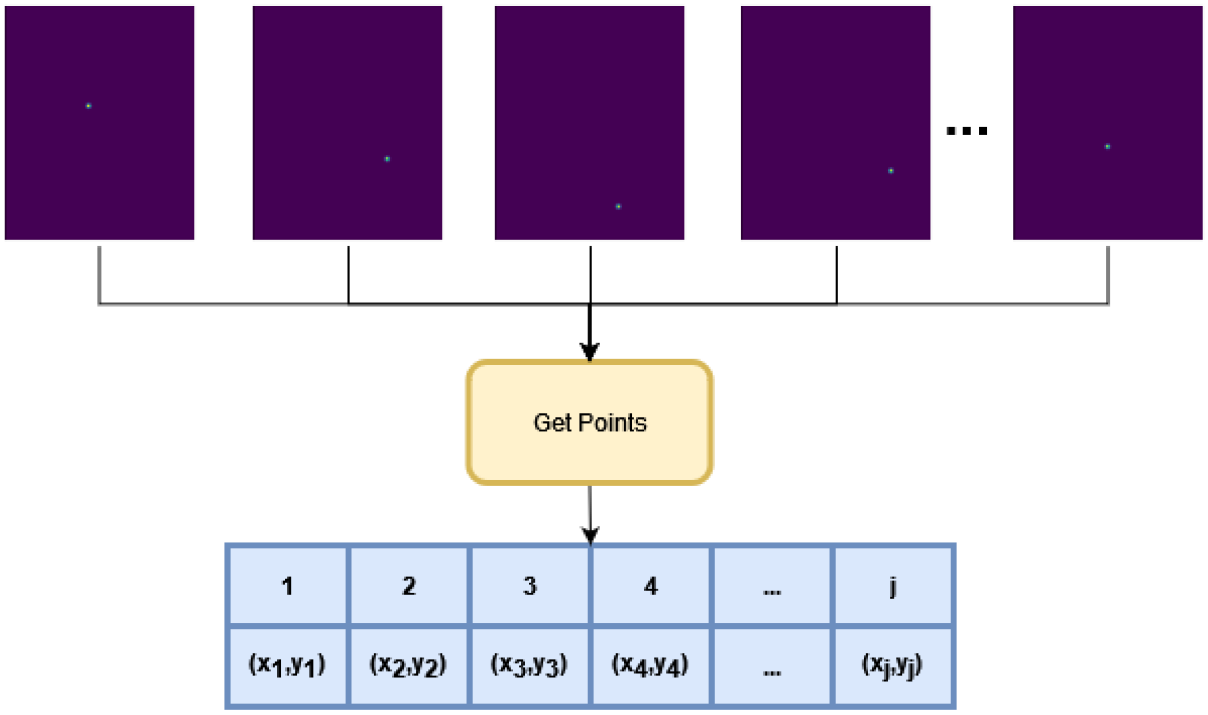


Figure 26 – Extraction of points in heat maps.

presented in Section 4.2.1 and 4.2.2. In the second proposal, the two models are fused, however, not all parts of the models are used. In both proposed models, fusion occurs during training between features from intermediate layers of the network and during testing by combining the outputs of two models.

4.3.1 Proposal I - Fusion with full models

We propose a model based on the fusion of two modules, M^p is the main model, such that $M^p(M^1, M^2)$. Module 1, M^1, is based on the model proposed by (ZHU et al., 2021; ZHU et al., 2022) and module 2, M^2, is based on the model proposed by (CHEN et al., 2019), as depicted by Figure 27.

Before fusing the model, we observed that in the feature extraction module proposed in (CHEN et al., 2019), the feature maps were created from VGG19 layers (Figure 21). From these observations, we concluded that the feature maps could be created using the resources of the layers of the down block proposed in (ZHU et al., 2021; ZHU et al., 2022). In this block, there are layers similar to those extracted from VGG19, as shown in Figure 28. Therefore, the feature extraction block and the down block form the link between the two models.

In addition to the creation of the link between the two models, it was necessary to adapt the entire (CHEN et al., 2019) proposal, so that it could act universally as proposed in (ZHU et al., 2021; ZHU et al., 2022). Module 2 was adapted to process images from

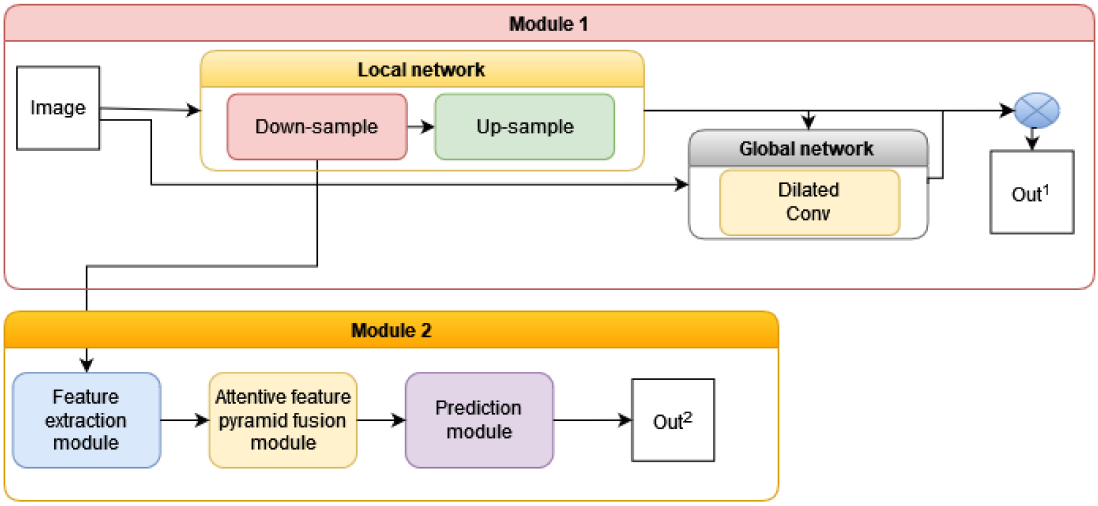


Figure 27 – Module 1 corresponds to the universal anatomical landmark detection model. Module 2 are the steps relating to the attentive feature pyramid fusion and regression-voting model. The fusion of modules occurs through the down-sample and feature extraction stage.

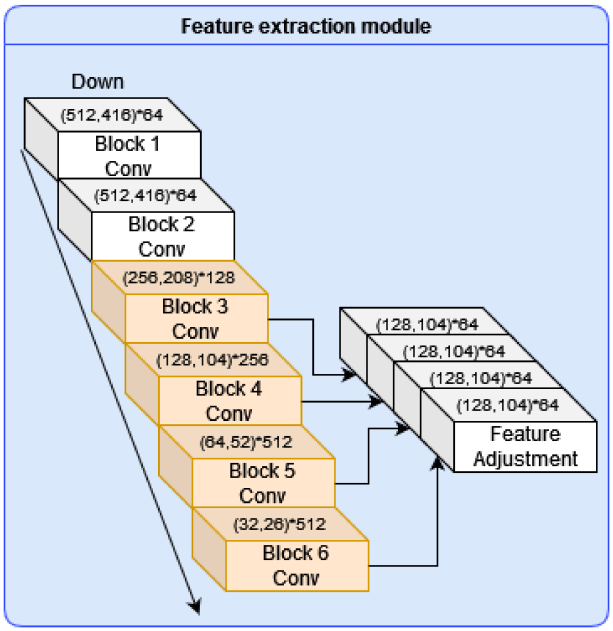


Figure 28 – The down convolution blocks are adapted for the feature extraction module. The input image is subjected to the first of six convolution blocks. The dimensions of the outputs of each convolution block are adjusted and concatenated.

different datasets and with different dimensions.

During the model training stage, the function to calculate the loss for each interaction is different for M^1 and M^2. Equation 13 is used to calculate the output loss of M^1 while Equation 11 for the loss of M^2. Using the PyTorch framework it was possible to define a retroactive function for both outputs. In this way, the model parameters were updated

recursively based on the responses of the two modules, Out^1 and Out^2. In the training process, the model instances used in the testing stage were saved in two stages.

The instance M_1^{best} was saved when the Out^1 output of M^1 reached the lowest loss value, and another instance M_2^{best} was saved when the response of Out^2 of M^2 reached the lowest loss value. During the testing phase, the model's final response was the fusion of the responses from the two saved models (M_1^{best} and M_2^{best}). This process is illustrated in Figure 29.

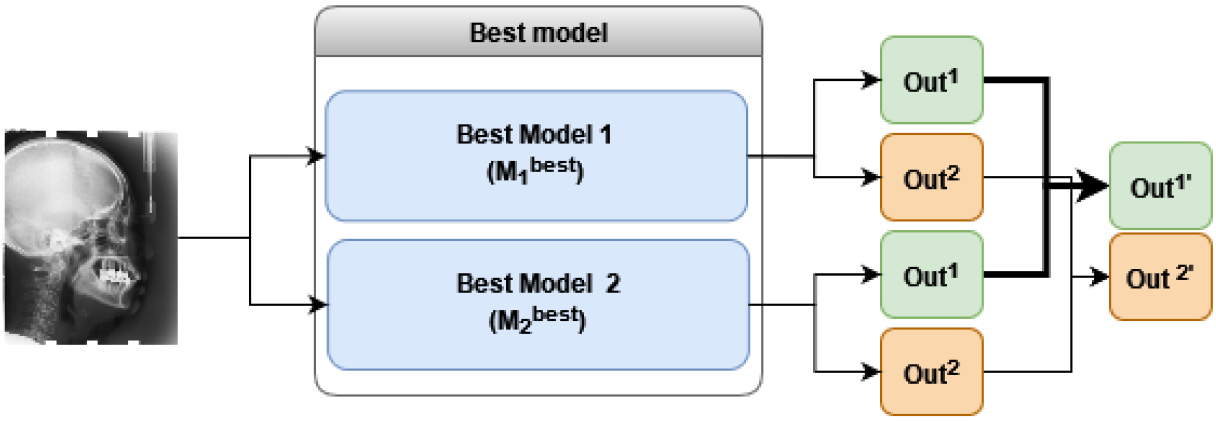


Figure 29 – Both best models generate two responses. The answers are combined to generate the final answer.

The output Out^1 is a tensor of dimensions (n, h, w) and output Out^2 is a tensor of dimensions $(n, 2)$; n are the landmarks, and $(h \times w)$ the dimensions of the input image. The output $Out^{1'}$ is the sum of the outputs Out^1 and the output $Out^{2'}$ is the sum of the outputs Out^2. The sum is performed using heat maps values normalized between 0 and 1.

4.3.2 Proposal II - Fusion with partial models

During the training of the model proposed by (CHEN et al., 2019), it was possible to observe the processing time of one dataset is similar to the processing time to process three datasets for the model proposed by (ZHU et al., 2021; ZHU et al., 2022). Through this observation, we consider proposing a second main model, U.

In this proposal, the models presented by (ZHU et al., 2021; ZHU et al., 2022) and (CHEN et al., 2019) are, respectively, Module 1, M^1, and Module 2, M^2, of U, such that $U(M^1, M^2)$. The U model is trained in two steps.

In the first stage, the heat maps of M^1 and M^2 are fused, as shown in Figure 30. In M^1 the local sub-module network L generates n heat maps in the form of a tensor T^l of size (n, h, w), n is the number of landmarks, h is the height of the input image and w is the width. The attentive feature Pyramid fusion sub-module $AFPF$, in M^2,

generates offset maps that we do not use in our proposal and heat maps which is a tensor T^p of size (n, h, w). A tensor T^f is obtained through the fusion of heat maps, such that $T^f = \alpha T^l + \gamma T^p$ (the sum of α and γ values is always equal to 1). The global network sub-module G, of the module M^1, processes the input image and the tensor obtained by fusion T^f and generates an output tensor T^g. The output F of the module M^1 is a combination of the tensors T^l and T^g, such that $F_k = T^g_k \odot T^l_k$, k is the number of images and the $\odot$ is pixel-wise multiplication, as presented on (ZHU et al., 2021; ZHU et al., 2022). The loss for M^1 output is calculated using Equation 13.

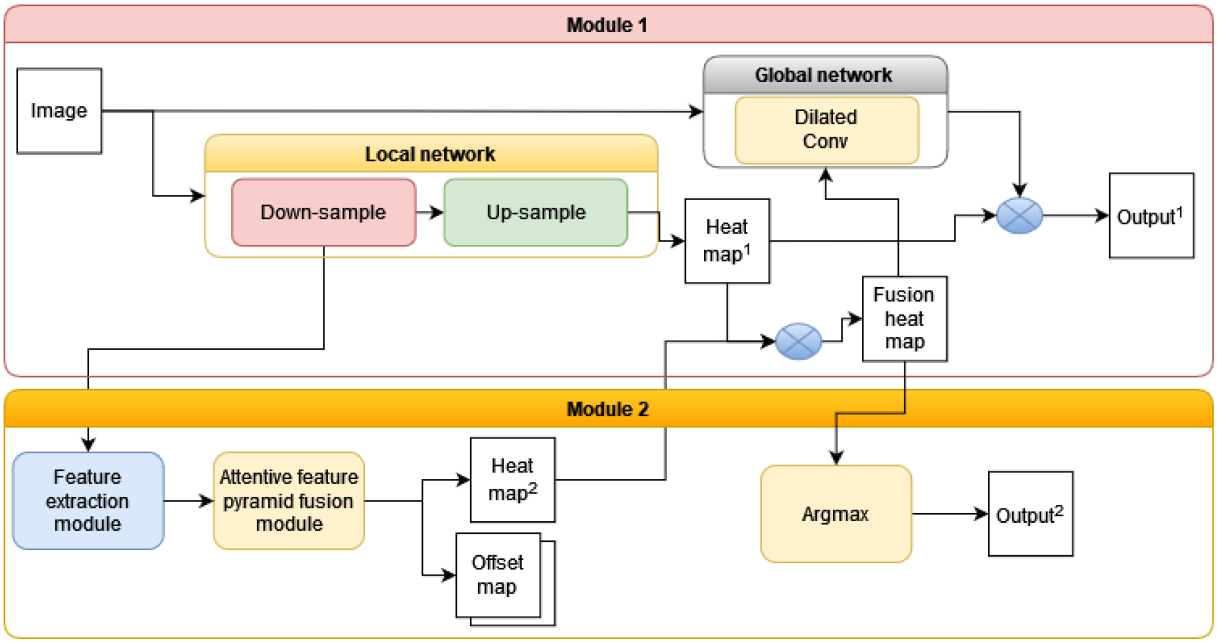


Figure 30 – Fusion between modules one and two. The modules are combined through the down-sample and feature extraction steps and through heat maps.

In the M^2 module, the heat maps obtained by fusion, tensor T^f, are processed by an *argmax* function that returns the position (x, y) in the n heat maps that have the largest value, resulting in an output tensor F^2 with size $(n, 2)$ – these are the prediction landmarks. The L2 loss is used to evaluate the response of the module M^2 (Equation 14). The final loss $L^f = (L^{M^1} + L^{M^2})/2$. At the end of the first training stage, the parameters of the model U that obtained the lowest loss, U_1^{best}, are stored.

$$L_k = \sum_{m \in C_k, n \in F_k^2} (c - p)^2, \tag{14}$$

C_k is a tensor of size (n,2) with the marked landmarks.

In the second stage, training starts from the U_1^{best} parameters and the model does not fuse the heat maps, as shown in Figure 31. The heat maps of M^2 are processed by the argmax function to generate the output tensor F^2. In addition, the heat maps are to the G sub-module of M^1. The final loss is calculated as follows: $L^f = (\upsilon L^{M^1} + \tau L^{M^2})/2$ (the

sum of υ and τ values is always equal to 1). After the second training stage the best model U, U_2^{best}, is stored. In the testing stage, the test dataset is submitted to the best-

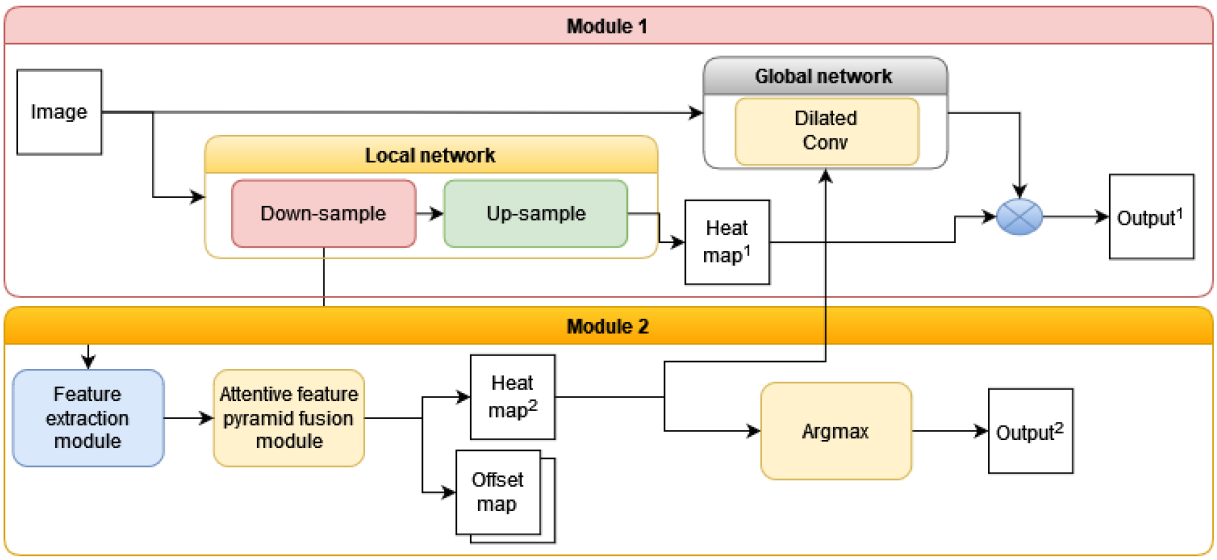


Figure 31 – Fusion between modules one and two. The modules are combined through the down-sample and feature extraction steps and the heat map from module two is submitted to the global network from module one.

stored models U_1^{best} and U_2^{best} and the final answer is a combination of the answer from M^1, such that $F_k^{pred} = F_k^{U_1} + F_k^{U_2}$. To obtain the final landmarks, an *argmax* function is used, $landmarks^{pred} = argmax(F_k^{pred})$, where $landmarks^{pred}$ is a tensor of size (n,2) (Figure 32).

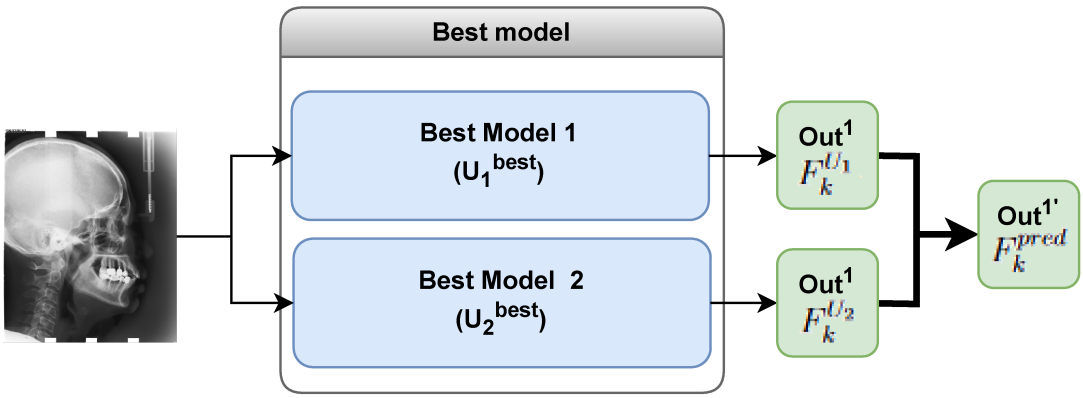


Figure 32 – Both best models generate one response. The answers are combined to generate the final answer.

4.3.3 Evaluation metrics

To evaluate the predictions of our models, we employ two metrics: the Mean Radial Error (MRE) and the Sucess Detection Rate (SDR). The MRE quantifies the Euclidean distance from the predicted reference point to the point manually marked by an expert. It is calculated using Equation 15 (ZENG et al., 2021):

$$MRE_j = \frac{\sum_{i=1}^{N} E_j^i}{N} \tag{15}$$

where N is the total number of images and the error is given by $E_j^i = ||\rho_j^p - \rho_j^m||_2$ for the image i, landmark j and $||.||_2$ represents the Euclidean norm function. Since $\rho_j^p = (x_j^p, y_j^p)$ and $\rho_j^m = (x_j^m, y_j^m)$, ρ_j^p are the predicted landmarks and ρ^m are the marked landmarks.

When dealing with reference points in medical images, there might be an acceptable margin of error in detecting these points. For instance, there could be scenarios where the system cannot deviate more than 2mm from the predicted point compared to the point marked by the expert. Therefore, to calculate the success detection rate for τmm, we use the following expression (ZENG et al., 2021):

$$SDR_\tau = \frac{count(\rho_j^p : ||\rho_j^p - \rho_j^m||_2) \leq \tau}{N} \tag{16}$$

The *count*() function is used to count how many reference points predicted by the model are at a distance smaller than τ mm from the point marked by the expert. Furthermore, the standard deviation (STD) was calculated using the *std*() method from the *numpy* library.

4.4 Results and Discussion

The experiments were carried out in the Google Colab environment, using a Tesla T4 GPU. Combining the datasets of lateral cephalograms, dataset containing radiographs of the hand and dataset containing radiographic images of the lung, there are 875 images for training, 113 for validation and 592 for testing.

As we propose, through fusion, an improvement in the universal model presented in Section 4.2.2, we use the same measurements used by the authors in the original work, MRE and SDR. For the lateral cephalograms dataset, the authors use four SDR measurements to present the results, 2mm, 2.5mm, 3mm, 4mm. For the dataset of radiographs of the hand, the SDR measurements to present the results were 2mm, 4mm, 10mm. In the dataset of radiographs of the lung, measurements of 3px, 6px, 9px were considered.

To analyze energy consumption and carbon emissions we consider the experiments carried out with the server running in Singapore. The average carbon emission per energy consumed (CO_2/kWh) for Singapore is 463.9. In the experiments, we report the estimated

energy consumption (kWh), equivalent carbon emissions (CO_2eq) and the equivalence of kilometers (Km) traveled by a car.

4.4.1 Experiments using Proposal I - Fusion with full models

In these experiments, the results obtained by running the model presented in Section 4.3.1 are presented. This model is used for landmark detection, it supports heterogeneous datasets with different numbers of landmarks. The new model combines characteristics of the universal model (Sections 4.2.1) and non-universal (Section 4.2.2). These models were run separately to check if the results coincided with the results presented in the theoretical basis. The universal model run for 100 epochs and the non-universal model run for 400 epochs. The results obtained through the execution of the precursor models, for the cephalometric images dataset, can be seen in the Table 8. After running the non-universal model (Section 4.2.1), it is estimated that model training uses 1.15 kWh of electricity contributing to 0.53 kg of CO_2eq. This is equivalent to 4.98 km traveled by car. For the universal model (Section 4.2.2), it is estimated that model training uses 0.66 kWh of electricity contributing to 0.30 kg of CO_2eq. This is equivalent to 2.84 km traveled by car.

Table 8 – Results of running the universal and non-universal model for the cephalometric image dataset.

Tests	Head				
	MRE	SDR(%)			
	(mm)	2mm	2.5mm	3mm	4mm
Chen et al. Section 4.2.1	1.35	81.03	87.77	92.06	96.57
Zhu et al. Section 4.2.2	1.49	76.99	84.08	89.62	94.93

The new universal model was trained four times over 50 epochs using four different random seeds. In Table 9 and 10, the results of the responses from the two modules of the new model are shown. Each "Run" present in the tables corresponds to a random seed. The various executions were to check whether the model presents stable results. Through the Tables, it is possible to see that at different times the model was executed, it showed little variation in results. The result of the first part (Table 9) is the best result of the CNN model created through fusion, for detecting landmarks in medical images. This result is used in comparison with the results of the literature.

After running the model proposed in this Section, it is estimated that model training uses 0.87 kWh of electricity contributing to 0.40 kg of CO_2eq. This is equivalent to 3.76 km traveled by car.

In Section 4.4.3 we compare the result of this proposal with the proposal by (ZHU et al., 2021; ZHU et al., 2022) and with our second proposal which is presented in Section 4.4.2.

Table 9 – Results referring to the response of Module 1 of the proposed model.

Tests	Head						Hand					Chest				
	MRE	STD	SDR(%)				MRE	STD	SDR(%)			MRE	STD	SDR(%)		
	(mm)	(mm)	2mm	2.5mm	3mm	4mm	(mm)	(mm)	2mm	4mm	10mm	(px)	(mm)	3px	6px	9px
Run 1	1.46	2.09	79.20	85.66	90.84	95.75	0.62	1.91	97.29	99.43	99.75	4.00	4.08	51.63	82.52	92.28
Run 2	1.46	1.89	78.80	85.16	89.89	95.37	0.62	1.83	97.29	99.43	99.69	4.12	4.57	49.59	81.71	93.50
Run 3	1.43	1.48	78.57	84.93	90.06	95.62	0.61	1.59	97.21	99.43	99.77	4.85	18.22	56.91	83.74	93.50
Run 4	1.43	1.48	79.62	85.24	90.40	95.31	0.57	1.32	97.37	99.57	99.85	4.36	7.53	54.47	81.71	89.02
Run average	1.45	1.74	79.05	85.25	90.30	95.51	0.60	1.67	97.29	99.46	99.76	4.33	8.60	53.15	82.42	92.07

Table 10 – Results referring to the response of Module 2 of the proposed model.

Tests	Head						Hand					Chest				
	MRE	STD	SDR(%)				MRE	STD	SDR(%)			MRE	STD	SDR(%)		
	(mm)	(mm)	2mm	2.5mm	3mm	4mm	(mm)	(mm)	2mm	4mm	10mm	(px)	(mm)	3px	6px	9px
Run 1	2.06	1.46	59.64	70.34	80.00	90.86	1.06	3.31	92.70	99.00	99.62	5.20	8.72	37.40	73.58	88.21
Run 2	2.03	1.43	60.88	71.79	80.65	91.41	1.05	2.95	92.01	99.00	99.60	5.23	7.72	38.21	70.73	86.99
Run 3	2.02	1.44	60.80	71.71	80.91	91.05	1.02	3.00	93.10	99.15	99.65	4.91	4.12	39.84	72.36	85.37
Run 4	2.02	1.40	60.46	72.15	81.33	91.28	1.03	2.83	92.76	99.13	99.66	5.37	7.69	35.77	72.76	87.40
Run average	2.03	1.43	60.45	71.49	80.72	91.15	1.04	3.02	92.64	99.07	99.63	5.18	7.06	37.80	72.36	86.99

4.4.2 Experiments using Proposal II - Fusion with partial models

This section presents the results obtained by running the model presented in Section 4.3.1. The proposed model is a fusion of the model presented in Section 4.2.1 with part of the model presented in Section 4.2.2. The new model detects landmarks in heterogeneous datasets and with different numbers of landmarks. The model is executed in two steps. In the first step, heat maps are merged, with $\alpha = 0.6536$, $\gamma = 0.3464$. In the second stage, the heat maps are not merged and weights are added to the loss combination, $\upsilon = 0.40$, $\tau = 0.60$. Both steps were performed for 10 epochs. Different values for the parameters α, γ, υ and τ were tested, however, this work presents the parameter settings for which we obtain the best results. The experiment was carried out 12 times with different random seeds, as shown in Table 11. The result is the average of the 12 experiments. In Table 12, 6 experiments are presented with $\upsilon = 0.30$ and $\tau = 0.70$.

Table 11 – Results of the new model with partial fusion ($\upsilon = 0.40$. $\tau = 0.60$).

Tests	Head						Hand					Chest				
	MRE	STD	SDR(%)				MRE	STD	SDR(%)			MRE	STD	SDR(%)		
	(mm)	(mm)	2mm	2.5mm	3mm	4mm	(mm)	(mm)	2mm	4mm	10mm	(px)	(mm)	3px	6px	9px
Run 1	1.40	1.36	79.71	85.98	90.72	95.85	0.55	0.99	97.67	99.60	99.93	4.02	4.27	55.28	82.11	89.84
Run 2	1.42	1.87	80.42	85.98	90.48	95.56	0.56	1.28	97.74	99.56	99.83	3.36	2.79	60.57	85.77	95.53
Run 3	1.43	1.28	78.53	84.63	89.58	95.12	0.57	1.16	97.47	99.66	99.87	3.97	3.42	49.59	82.93	93.90
Run 4	1.42	1.41	79.56	85.68	90.42	95.68	0.55	1.25	97.70	99.62	99.91	3.65	3.33	55.28	83.74	93.09
Run 5	1.41	1.28	79.68	85.73	90.63	95.77	0.57	1.18	97.72	99.56	99.85	3.65	3.92	60.16	86.99	94.72
Run 6	1.44	1.36	79.35	85.41	90.25	96.00	0.57	1.27	97.75	99.53	99.87	3.43	3.29	57.72	85.77	94.72
Run 7	1.41	1.43	79.66	86.17	90.59	95.75	0.58	1.54	97.63	99.52	99.85	3.90	3.92	54.88	82.93	89.84
Run 8	1.44	1.37	79.09	85.47	89.94	95.14	0.76	3.34	97.16	99.17	99.45	3.87	3.63	57.32	81.30	92.28
Run 9	1.43	1.30	79.03	85.43	90.44	95.71	0.56	1.24	97.52	99.67	99.90	3.30	2.66	56.10	89.84	95.93
Run 10	1.41	1.87	80.38	86.46	91.18	96.34	0.54	0.72	97.65	99.67	99.96	3.57	3.16	60.16	82.93	92.28
Run 11	1.37	1.20	80.78	86.93	91.39	96.29	0.57	1.35	97.47	99.56	99.85	3.92	3.36	50.00	81.71	91.06
Run 12	1.41	1.31	80.23	85.85	90.29	95.49	0.57	1.46	97.64	99.60	99.87	3.60	3.21	55.69	84.55	92.68
Run average	1.41	1.42	79.70	85.81	90.49	95.72	0.58	1.40	97.59	99.56	99.84	3.69	3.41	56.06	84.21	92.99

Table 12 – Results of the new model with partial fusion ($\upsilon = 0.30$, $\tau = 0.70$).

Tests	Head						Hand					Chest				
	MRE	STD	SDR(%)				MRE	STD	SDR(%)			MRE	STD	SDR(%)		
	(mm)	(mm)	2mm	2,5mm	3mm	4mm	(mm)	(mm)	2mm	4mm	10mm	(px)	(mm)	3px	6px	9px
Run 1	1,39	1,24	80,40	86,23	91,14	96,04	0,54	0,93	97,69	99,62	99,94	4,05	4,10	55,69	80,08	90,24
Run 2	1,42	2,38	80,55	86,29	90,61	95,85	0,58	1,52	97,83	99,54	99,79	4,31	13,17	59,76	83,74	95,53
Run 3	1,45	1,49	78,74	84,82	89,41	95,18	0,57	1,18	97,45	99,61	99,88	3,81	3,22	50,81	83,74	94,31
Run 4	1,44	1,71	79,09	85,33	90,27	95,62	0,56	1,26	97,67	99,62	99,90	3,69	3,29	55,69	81,30	93,50
Run 5	1,45	1,39	79,39	85,52	90,15	95,16	0,56	1,04	97,71	99,60	99,87	3,64	3,32	56,91	84,55	93,50
Run 6	1,43	1,28	78,97	85,20	89,94	95,89	0,59	1,37	97,54	99,49	99,79	3,37	2,78	56,50	85,37	93,90
Run average	1,43	1,58	79,52	85,56	90,25	95,62	0,57	1,22	97,65	99,58	99,86	3,81	4,98	55,89	83,13	93,50

After running the model proposed in this Section, it is estimated that model training uses 0.23 kWh of electricity contributing to 0.11 kg of CO_2eq. This is equivalent to 0.98 km traveled by car.

In Section 4.4.3, the results presented in this section and the results of Proposal I are compared to the results presented in the literature.

4.4.3 Discussion

In this section, we compare the results of our proposals with results from the literature. Table 13 shows that, our proposals in the last two lines of the table, on average, presented better results than those presented in the literature.

Table 13 – Evaluation metrics – * indicates that the values were obtained in the cited papers. In bold are the best results. Underlined are the second best results.

Studies	**Head**					**Hand**				**Chest**			
	MRE	**SDR(%)**				**MRE**	**SDR(%)**			**MRE**	**SDR(%)**		
	(mm)	**2mm**	**2.5mm**	**3mm**	**4mm**	**(mm)**	**2mm**	**4mm**	**10mm**	**(px)**	**3px**	**6px**	**9px**
U-Net*	12.45	52.08	60.04	66.54	73.68	6.14	81.16	92.46	93.76	5.61	51.67	82.33	90.67
GU2Net*	1.54	77.79	84.65	89.41	94.93	0.84	95.40	99.35	99.75	5.57	**57.33**	82.67	89.33
Fusion with full models (Ours)	1.45	79.05	85.25	90.30	95.51	0.60	97.29	99.46	99.76	4.33	53.15	82.42	92.07
Fusion with partial models (Ours)	**1.41**	**79.70**	**85.81**	**90.49**	**95.72**	**0.58**	**97.59**	**99.56**	**99.84**	**3.69**	56.06	**84.21**	**92.99**

We observed that Proposal I, compared to the model proposed by (ZHU et al., 2021; ZHU et al., 2022), has better results, however, it takes longer to execute 50 epochs during the training stage. The increase in time occurs because in the fusion of models, at each iteration, the loss is calculated at five different moments. A loss is calculated for Module 1 heat maps, Module 2 heat maps, and the two Module 2 displacement maps. Additionally, the GPU RAM consumption increases.

In the experiment related to Proposal II, we observed that the execution time of the new model is slightly longer than previous proposals (ZHU et al., 2021; ZHU et al., 2022) because the loss is calculated at two different times. However, the proposed model achieves good results early. To achieve the best results, our model ran for 20 epochs, while the literature proposal ran for 100 epochs. After the experiments, Proposal II proved to be the best solution.

Figure 33 shows the performance by landmark point detection for all dataset. The comparison is performed between the original model and the Proposal II model. It is

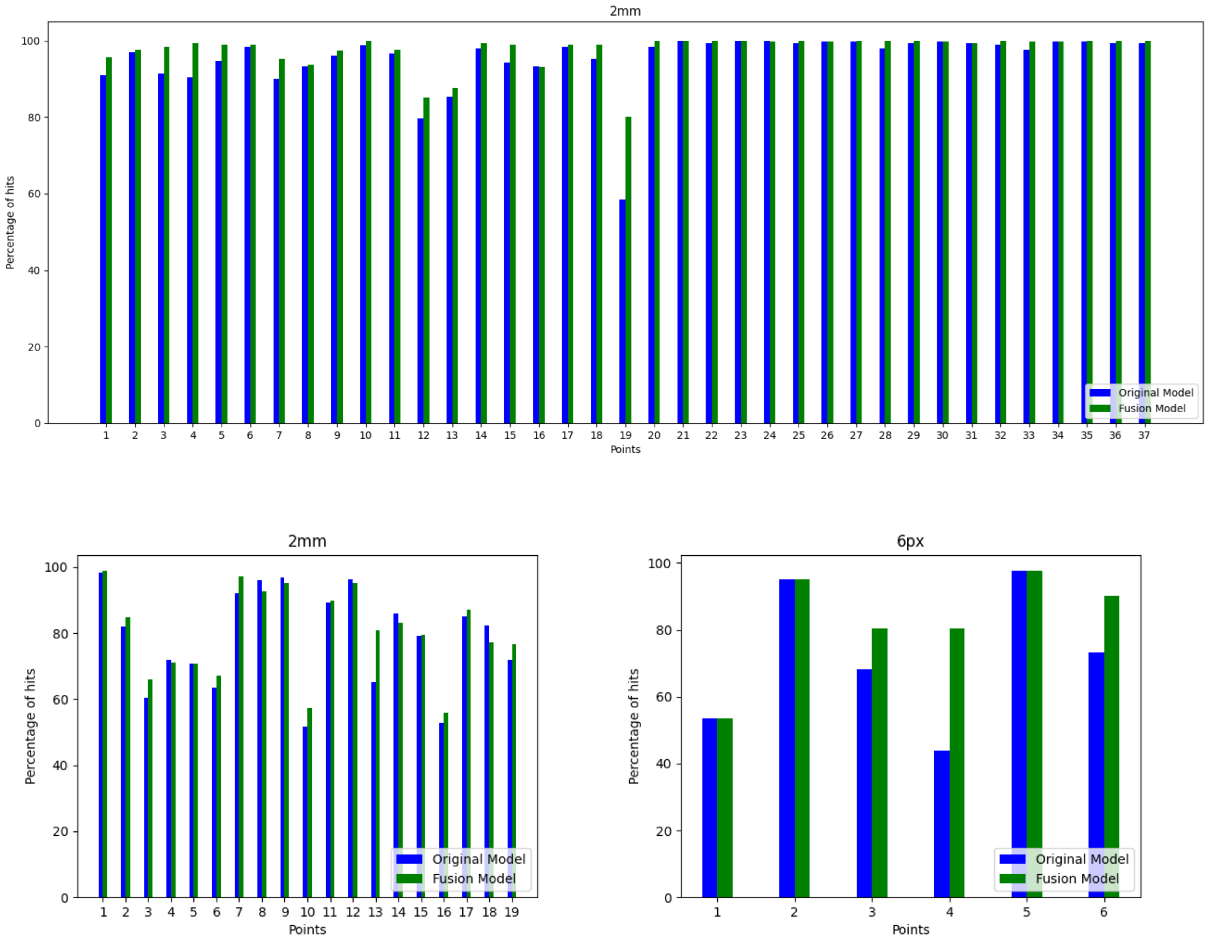


Figure 33 – From top to bottom: the first graph shows the model's performance for each point in the hand radiography dataset. The graph on the left shows the performance, per point, for the lateral cephalogram dataset. The graph on the right shows the performance for the lung X-ray dataset.

observed that the Proposal II model, created by fusion, achieved better results in detecting multiple points.

Analyzing energy consumption and carbon dioxide emissions, the non-universal model presented the worst performance, needed to be run for 400 epochs, just for the lateral cephalogram dataset, spent 11.41 hours, consumed 1.15 kWh and emitted 0.53 kg of $CO2eq$. The universal model was run for 100 epochs, for a mix of three datasets (lateral cephalograms, hand radiographs and lung radiographic), spent 7.52 hours, consumed 0.66 kWh and emitted 0.30 kg of $CO2eq$. Proposal II, a fusion between the universal and non-universal model, obtained the best performance. It spent 2.31 hours, consumed 0.23 kWh and emitted 0.11 kg of $CO2eq$, reducing execution time, energy consumption and carbon emissions by around 65%. Furthermore, this proposal used a mix of three datasets. In Figure 34 it is possible to visualize the CO_2eq for the models trained and presented in this chapter. The graph shows that the non-universal model is the one

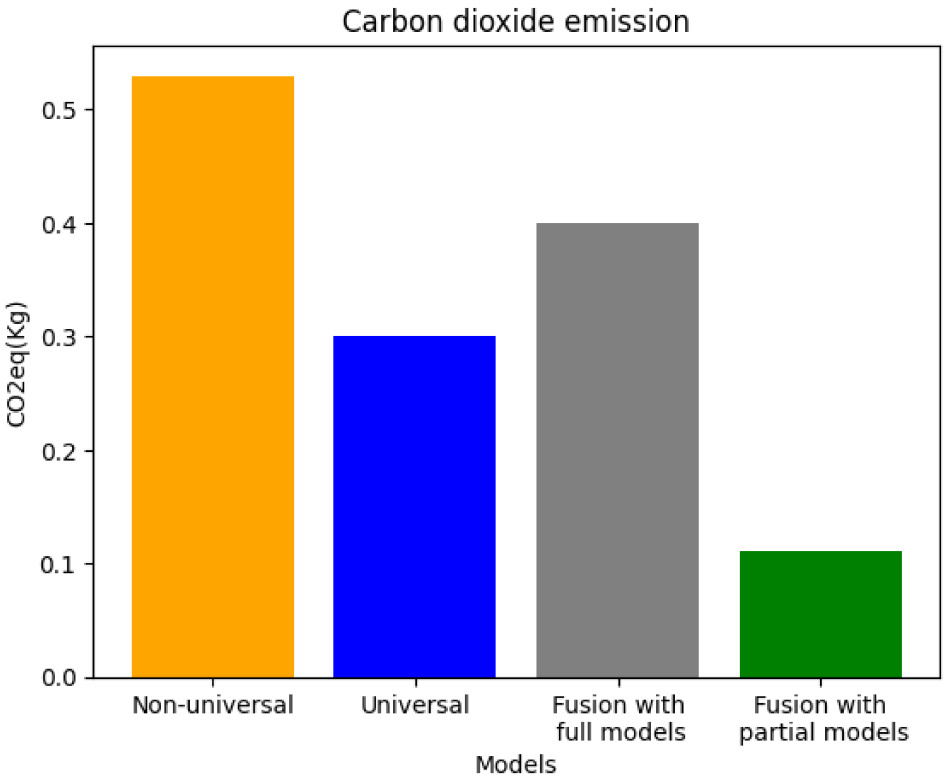


Figure 34 – Estimated carbon emission for each model.

that consumes the most resources and our proposal for partial fusion between the models consumes less resources to achieve the results in the literature.

Chapter 5

Influence of preprocessed images on the performance of CNNs

In image analysis, a common task is preprocessing. This step is essential for the optimal performance of many techniques. In this chapter, we propose a preprocessing method based on histogram equalization and Benford's distribution. Histogram equalization aims to enhance image contrast and quality by redistributing pixel intensities. The application of Benford's law involves presegmentation of the image based on statistical distributions observed in the leading digits of pixel values.

Additionally, we conducted an analysis to determine whether CNNs used for landmark detection or region segmentation perform better when evaluated on datasets with preprocessed images. We check whether models trained on the original dataset and the dataset with preprocessed images, when fused, achieve better performance. In alignment with green computing, we also assessed the resource consumption of the models.

5.1 Background

This section provides an insight into image preprocessing, histogram equalization, fundamentals of Benford's law and the hybridGNet model used for region segmentation.

5.1.1 Image preprocessing

The high capacity of convolutional neural networks allows them to process images with their original values, dimensions, and pixels. However, in some cases, the performance of convolutional networks can be improved with the use of suitable preprocessing methods. Authors report in articles the influence of preprocessing on the performance of CNNs (TABIK et al., 2017; ÖZTÜRK; AKDEMIR, 2018; AFIFI; BROWN, 2019; LIN; CHANG, 2021; ARABIAN et al., 2021; GIEŁCZYK et al., 2022).

In convolutional neural networks, a common preprocessing process is data augmentation. This process consists of applying different transformations to the training and validation set, seeking to reduce overfitting and noise sensitivity. However, other traditional preprocessing techniques can be used, such as mean filter, median filter, Sobel filter, and histogram equalization, among others. These techniques allow you to highlight important data or remove irrelevant data from the image before it is processed by CNN (SONKA et al., 1993; TABIK et al., 2017; ÖZTÜRK; AKDEMIR, 2018; AFIFI; BROWN, 2019; LIN; CHANG, 2021; ARABIAN et al., 2021; GIEŁCZYK et al., 2022).

In this chapter, experiments are performed that seek to improve the detection of regions in images using CNNs. Histogram equalization and Benford's law techniques are combined to perform the preprocessing of the images analyzed by the CNNs.

5.1.2 Benford's law

Benford's law, known as the law of the first digit, was observed by Simon Newcomb in 1881, but became famous through the publications of Frank Benford in 1938. Later, Theodore P. Hill proved that this law is universally applicable, being invariant to scale and base. Benford's law deals with the statistics of natural phenomena. In a set of numbers that quantify a natural phenomenon, it is common to observe Benford's law. This law says that, considering the digits from 1 to 9, the probability that the digit 1 appears as the first digit of a number, in a set of numbers that represents a natural phenomenon, is greater than the probability that the digit 2 appears as the first digit of a number. The probability of digit 2 is greater than the probability of digit 3, and so on, up to digit 9, as shown in Figure 35 (BENFORD, 1938; HILL, 1995; ACEBO; SBERT, 2005;

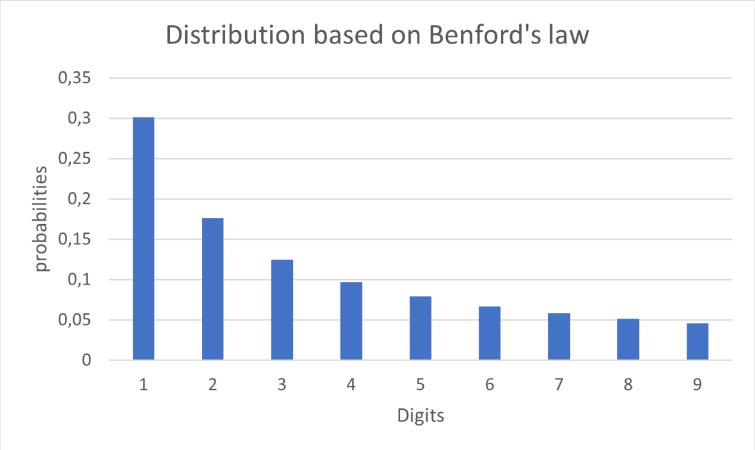


Figure 35 – Logarithmic distribution of digits using Benford's law. Adapted from: (BENFORD, 1938)

SAMBRIDGE; TKALČIĆ; JACKSON, 2010; BERGER; HILL, 2011).

By means of Benford's law, a logarithmic distribution is observed for the most significant digit of a number, given by the Equation 17 (BERGER; HILL, 2011).

$$Prob(D_1 = d_1) = log_{10}(1 + d_1^{-1}) \tag{17}$$

for $d_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and D_1 is the most significant digit of a number.

In the literature, Benford's law is commonly used to detect fraud in accounting data (DURTSCHI et al., 2004). However, in the area of image processing, some authors use Benford's law in different situations. In (JOLION, 2001) and (ACEBO; SBERT, 2005), situations are shown in which images do or do not present the characteristics of Benford's law. In (MAKRUSHIN et al., 2018) and (PARNAK; DAMAVANDI; KAZEMITABAR, 2022), the authors use Benford's law to detect image fraud. In (WELLS et al., 2007) and (AL-BANDAWI; DENG, 2019), the law is used in image distortion analysis. In (AL-BANDAWI; DENG, 2019), the authors even use Benford's law in preprocessing. The first digit frequency information is used to compose a feature vector in a system that classifies distortions in images. In this work, the law is used as a preprocessing alternative.

5.1.3 Histogram equalization

Histogram equalization is a contrast enhancement technique that seeks to distribute pixel values within the range of image intensity. If the intensity range of an image corresponds to all integer values between 0 and 255, but the image consists of only 40% of the values in this range, histogram equalization seeks to generate a new image in which the percentage of values in the range from 0 to 255 tends to 100%. The histogram equalization is expressed by Equation 18 (PIZER et al., 1987; GONZALEZ; WOODS, 2009; ABDULLAH-AL-WADUD et al., 2007).

$$s_k = T(r_k) = (L-1)\sum_{j=0}^{k} p_r(r_j) = \frac{(L-1)}{MN}\sum_{j=0}^{k} n_j \tag{18}$$

Being $L-1$ the value of 255, $k = 0, 1, 2, ..., L-1$, r_k the value of the pixel before transformation, s_k the value output pixel after transformation, $T(r_k)$ is the histogram equalization transform, M and N the image dimensions, n_j is the image pixel quantity with the intensity (r_j and $p_r(r_j)$ is the probability estimate that r_j occurs in the image (GONZALEZ; WOODS, 2009). Histogram equalization is a commonly used technique for preprocessing images (GONZALEZ; WOODS, 2009; SHIN; KIM; KWON, 2016; ALWAWI; ABOOD, 2021).

5.1.4 HybridGNet for region detection

HybridGNet[1] is an architecture for region segmentation based on landmarks in medical chest x-ray images. In (GAGGION et al., 2021; GAGGION et al., 2022) the authors demonstrate the performance of the network to segment the lung, heart and clavicle regions. HybridGNet combines traditional convolutional neural networks with graph convolutional neural networks to increase the accuracy of segmenting anatomical structures, Figure 36. By integrating graph convolutional neural networks, the model can better capture the complex relationships between anatomical landmarks, leading to more accurate segmentation results (GAGGION et al., 2021; GAGGION et al., 2022).

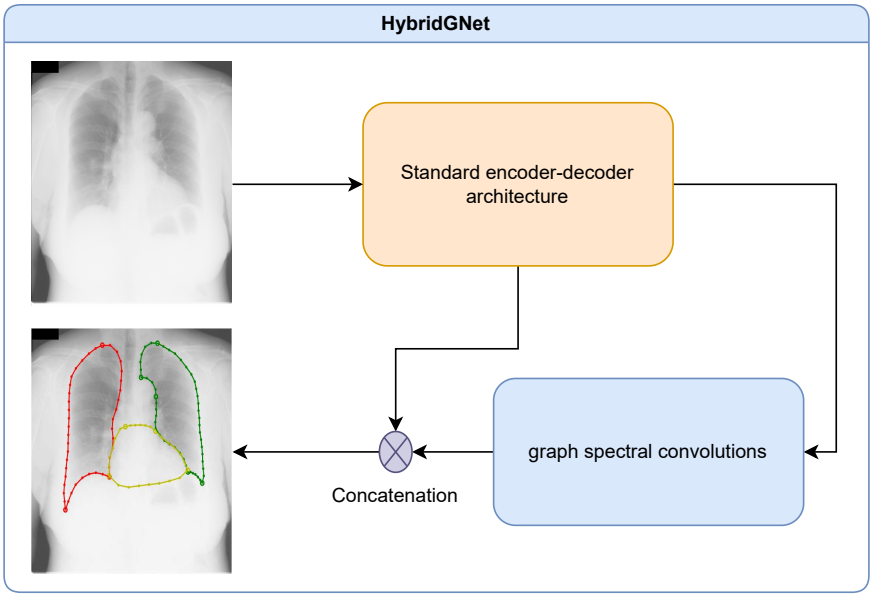


Figure 36 – The HibridGNet model combines the features of a standard convolutions architecture with graph spectral convolutions.

The model's hybrid approach leverages local and global image information, enabling detection of anatomical details not found in conventional models. As described (GAGGION et al., 2021; GAGGION et al., 2022), HybridGNet performed better than traditional landmark based models, making it reliable for segmenting anatomical regions in medical applications (GAGGION et al., 2021; GAGGION et al., 2022).

5.2 Proposal

The first proposal is to preprocess datasets through histogram equalization and a method based on Benford's law. The second proposal is to use CNNs to process images

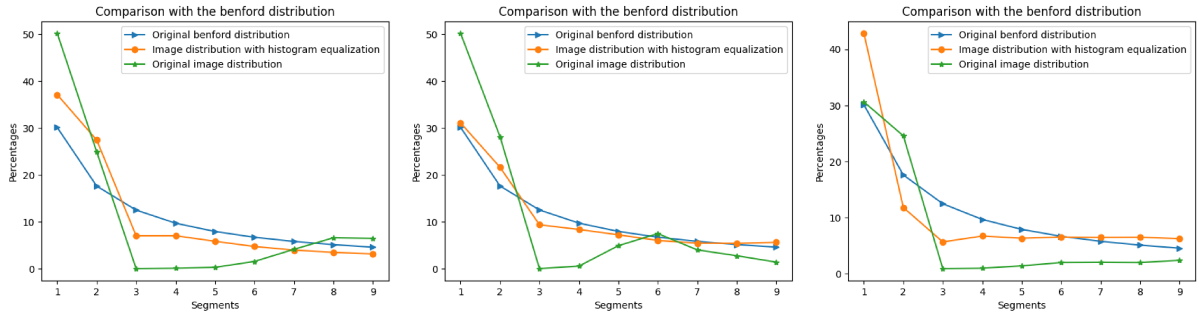
[1] Available at: <https://github.com/ngaggion/HybridGNet>

from the original datasets and modified datasets and, in addition, perform the fusion of models trained with the different datasets.

5.2.1 Proposal I - Creating image dataset using a method based on Benford's law

In this chapter we call the proposed method the Benford method. It is applied to images in gray levels, with pixel values between 0 and 255. Before applying the Benford method, the first action was to equalize the histogram. This procedure helped to balance the gray levels of the image, preventing the gray levels from being concentrated in a few values.

After applying histogram equalization to the images in the original dataset, through Figure 37 it is possible to observe how much the histogram equalization brings the distri-



(a) Graph for the dataset with images of lateral cephalograms.

(b) Graph for the dataset with radiograph images of the hand.

(c) Graph for the dataset with radiographic images of the lung.

Figure 37 – Graphs that show the approximation of the Benford curve, in each dataset, after the histogram equalization. In the graphs, the blue line with the symbol ▶ indicates the original Benford distribution (literature standard). The green line with the ⋆ symbol shows the original distribution of the dataset. The orange line with the • symbol shows the distribution of the dataset after the images have gone through histogram equalization.

bution of pixels in the original images closer to the Benford distribution. Observing the green line in Figures 37(a), 37(b), 37(c), it is noticeable that the distribution of pixels in original images (green line) were far from the Benford distribution (blue line). After equalizing the histogram, data represented by the orange line, the distribution of pixels in the images in the dataset was close to the Benford distribution (blue line).

In Figure 38 it is possible to view the original image and the same image after going through histogram equalization. After equalizing the histogram, regions of the images were extracted based on Benford's law.

For image processing based on Benford's law, consider a dataset of images $A_{i,j}$ with x pixel values that vary between 0 and 255, Equation 19.

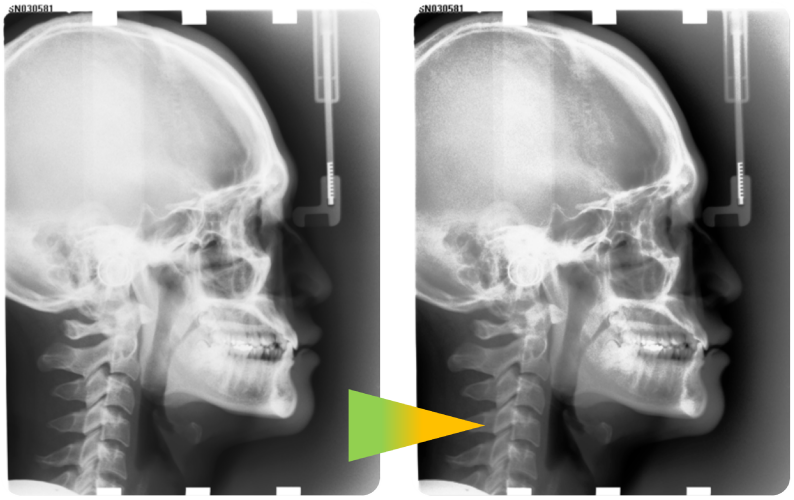


Figure 38 – From left to right: the first is the original image and the second is the original image after equalizing the histogram.

$$A_{i,j} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{19}$$

For $\forall x_{ij} \in A_{ij}$ has been, Equação 20:

$$f(x) = \begin{cases} \lfloor \frac{x_{ij}}{10} \rfloor, & x_{ij} \in [10, 99] \\ \\ \lfloor \frac{x_{ij}}{10^2} \rfloor, & x_{ij} \in [100, 255] \end{cases} \tag{20}$$

Now consider an intermediate matrix that receives the result of $f(x)$, $A'_{ij} = f(x)$, considering that $d \in \{1, 2, \cdots, 9\}$, for all d the function $g(x)$, Equation 21, generates a binary image when applied a $f(x)$, tal que $S^d_{ij} = g(f(x))$.

$$g(x) = \begin{cases} 1, & x_{ij} \in A'_{ij} = d \\ 0, & x_{ij} \in A'_{ij} \neq d \end{cases} \tag{21}$$

At the end of the process, for each $A_{i,j}$ image, nine images $\{S^1, S^2, ..., S^d\}$ are generated, these images are called segments, as shown in Figure 39. Resulting in the finite union of $d = 9$ segmented images, described in Equation 22.

$$S'_{ij} = \bigcup_{d=1}^{9} S^d_{ij} \tag{22}$$

Some parts generated from the image with the equalized histogram were selected, as shown in Figure 41. The selection of images considered Benford's law, which shows that approximately 40% of the values start with the digits 1 and 2. Thus, images 1 and 2 were selected. The images representing the other digits contained contour information.

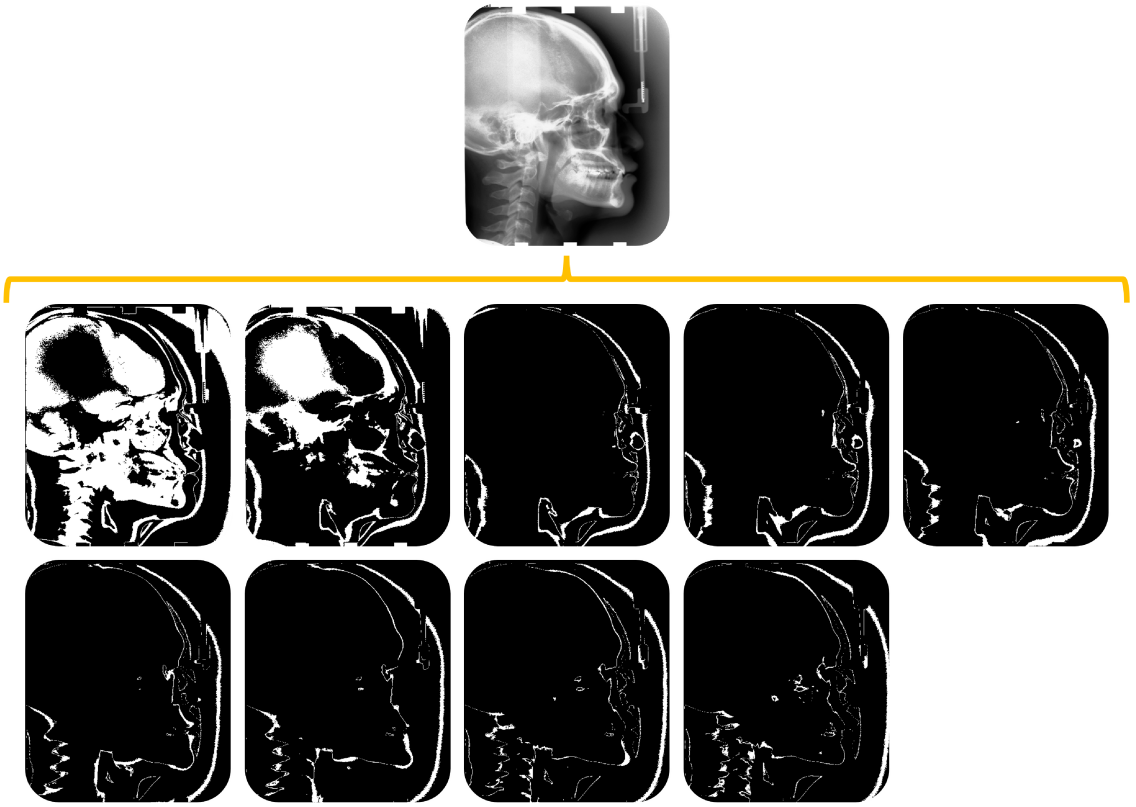


Figure 39 – Image equalization and bitplane images. From top to bottom: the first is the original image after equalizing the histogram. In the first sequence of images, from left to right, the first image shows all the pixels that start with the digit 1 (S_1), the second one refers to all the pixels that start with the digit 2 (S_2), and so on, until the last image of the second sequence , in which all pixels starting with the digit 9 (S_9) are represented.

Contour details and stroke thickness were considered criteria for selecting another image. The image showing pixels starting with the digit 9 had thin lines; however, it was observed that in some situations there was a loss of contour information. The image containing the pixels starting with the digit 8 was chosen, as it presented many contour details and fine lines. The image selection process for digit 8 is manual based on data observation. Therefore, images referring to the digits 1 (S_1), 2 (S_2) and 8 (S_8) were selected.

The selected images were joined using a sum operation, resulting in the image shown in Figure 41. This image was transformed into a binary image, where the white pixels have the value 1, and the black pixels have the value 0.

The binary image was multiplied by the original image, which underwent histogram equalization, generating a new image, as shown in Figure 42. The new image highlighted only the regions where the pixels had values starting with the digits 1, 2 and 8.

Each original image went through this sequence of procedures, generating a new set of data, as shown in Figure 43.

Figure 40 – From left to right: in the first sequence of images, the first and second images were selected, which present the pixels that start with the digit 1 (S_1) and 2 (S_2), respectively. In the second sequence of images, the third image is selected, which presents all pixels that begin with the digit 8 (S_8).

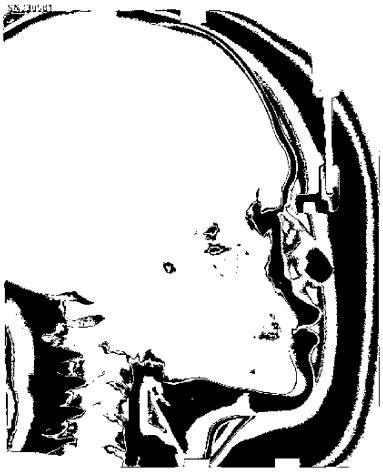


Figure 41 – This image shows the sum result of images in which the pixels starting with the digits 1 (S_1), 2 (S_2) and 8 (S_8) are white.

5.2.2 Proposal II - Datasets with Benford images and CNN for landmark detection

The dataset with images generated based on Benford's law and the original dataset were submitted, for training, to a CNN model proposed by (ZHU et al., 2021; ZHU et al., 2022) and presented in Section 4.2.2. At the end of training, the best models are stored, those in which the parameters produced the best results, as shown in Figure 44.

During the testing stage, the model responses are fused. The heatmaps generated from the original datasets were fused with the heatmaps generated from the Benford datasets via a summation operation. From the heat maps resulting from the sum, the vector of points was generated, as shown in Figure 45.

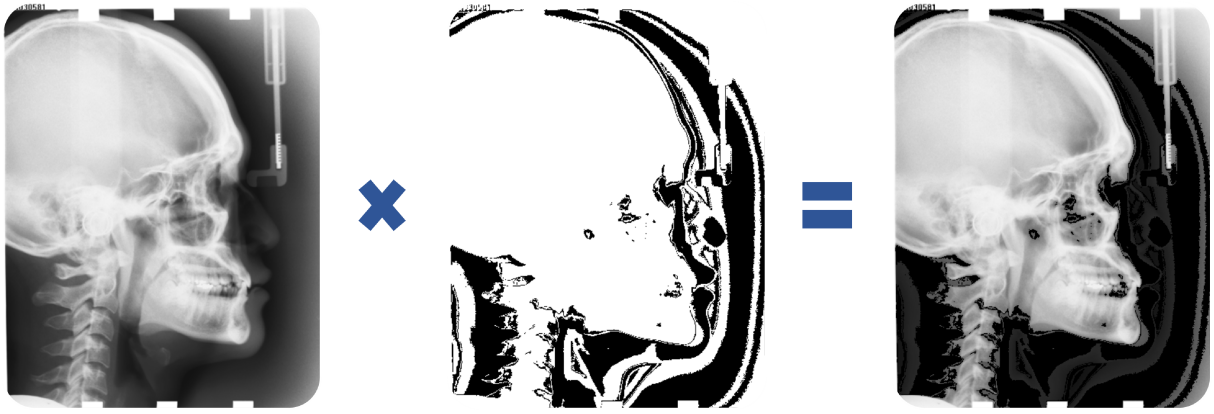


Figure 42 – The original image is multiplied by the image resulting from the sum of selected images.

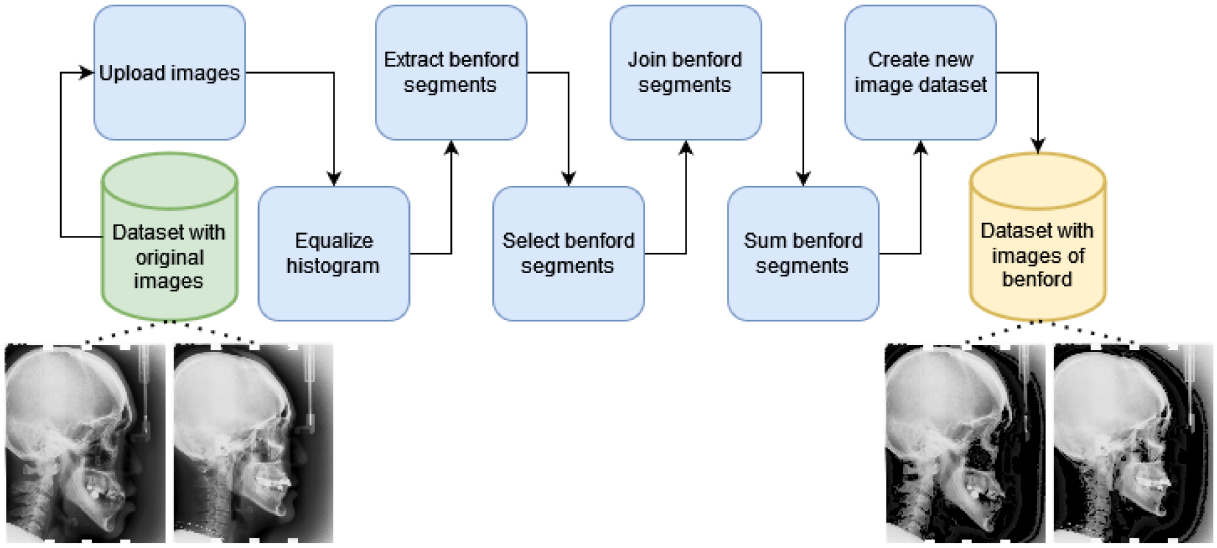


Figure 43 – Steps of the method based on Benford's law.

In Figure 46 the process of fusing the responses of the CNNs is represented. The original image contains the original marks (ground truth). To the left of the original image, in the first image, there are the landmarks detected from the original dataset image. The second image is the comparison of ground truth with the landmarks detected from the original dataset image. To the right of the original image, in the first image, there are the landmarks detected from the Benford dataset image. In the second image is the comparison of ground truth with the landmarks detected from the Benford dataset image. In the center, below the original image, are the markings resulting from the fusion. The third central image is the comparison of ground truth with the landmarks obtained through the fusion. The fusion is the sum of the landmarks obtained based on the original

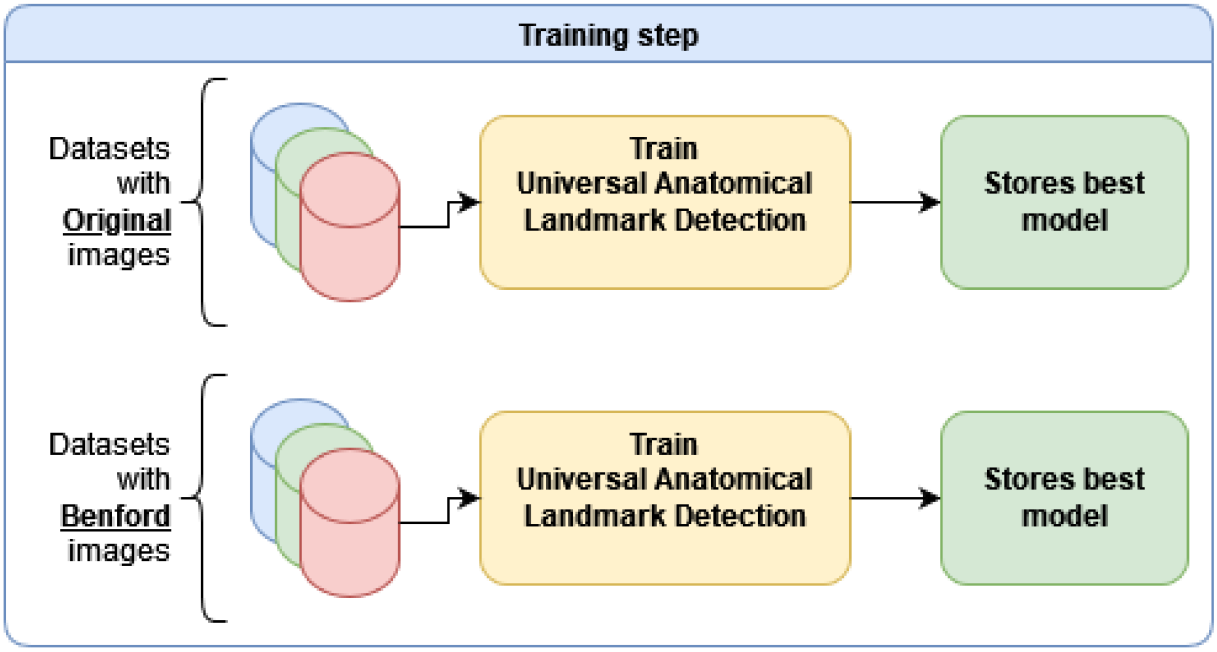


Figure 44 – Training the universal anatomical landmark detection model with the original dataset and the Benford dataset.

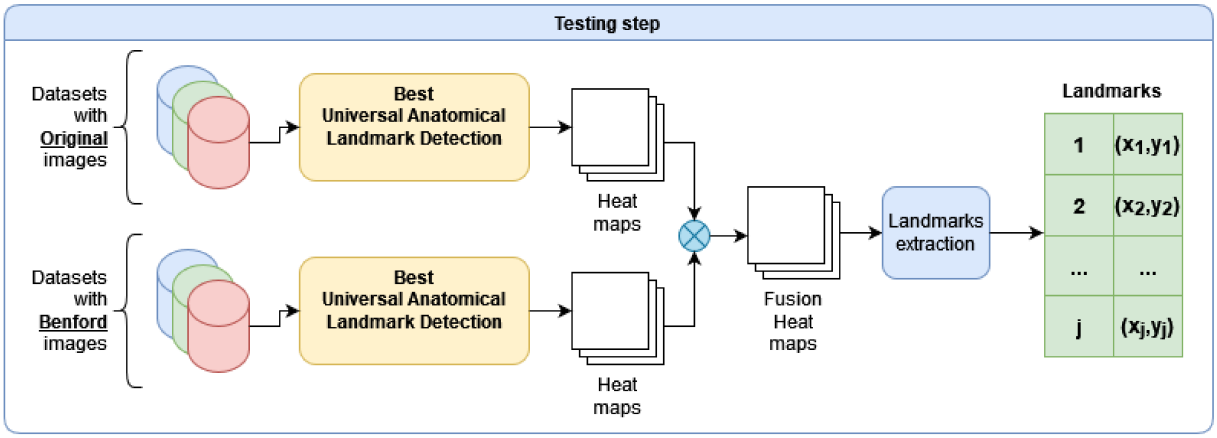


Figure 45 – Fusion process of the universal model trained with the original dataset and the model trained with the Benford dataset. The output of the models are heat maps and the fusion is the sum of the heat maps. The landmarks are extracted from the summed heat maps.

dataset and the landmarks obtained with the Benford dataset.

When using HybridGNet, Section 5.1.4, the process during the training phase is the same as shown in Figure 44. The model is trained and at the end of training the best models are stored. In the testing phase, the responses for the models trained on the two datasets are averaged, as shown in Figure 47.

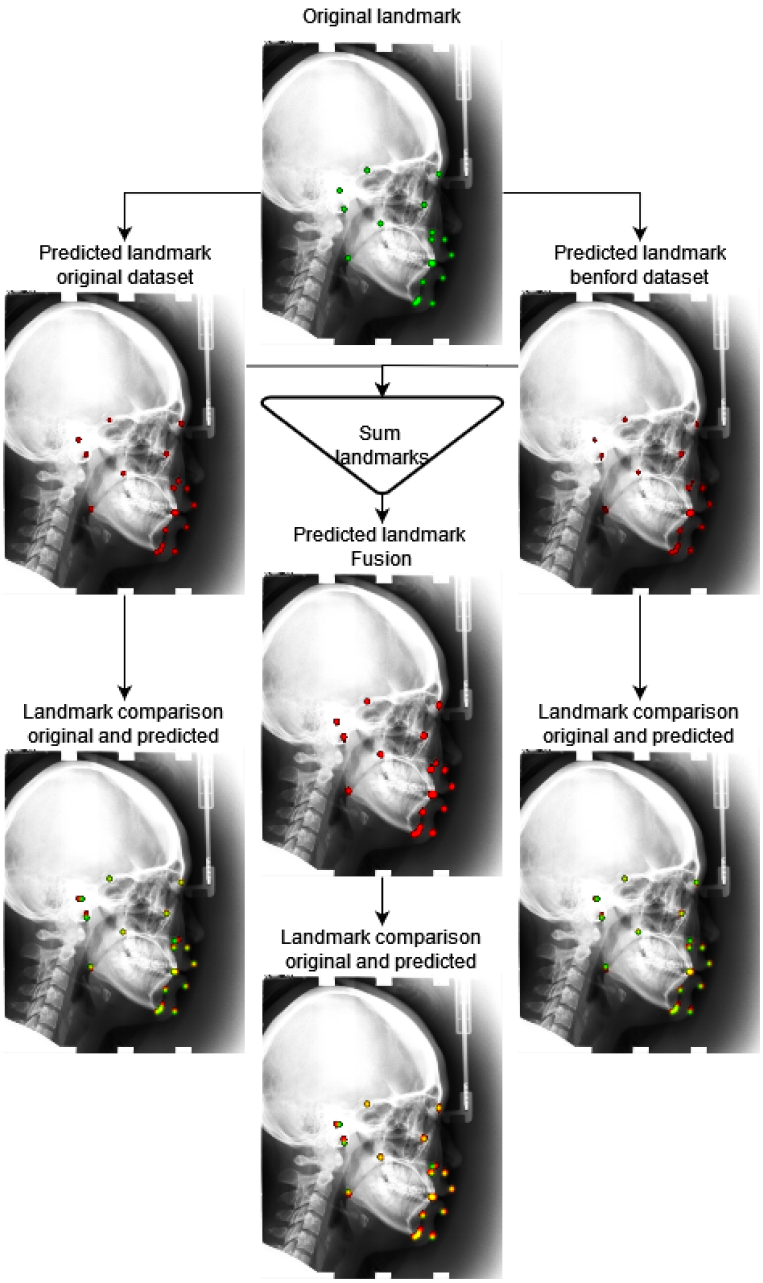


Figure 46 – Process of fusion the answers from CNNs.

5.3 Results and discussion

The experiments presented in this section evaluate the contribution of datasets created using the Benford method for landmark detection and region segmentation. In experiments involving landmark detection the universal model, Section 4.2.2, is used. The HybridGNet architecture, Section 5.1.4, is used in experiments involving region segmentation based on the relationship between anatomical landmarks.

In the experiments, four datasets with the original images and four datasets with the images processed by the Benford method is used, as shown in Figure 48.

For the experiments involving the universal model, the datasets of lateral cephalo-

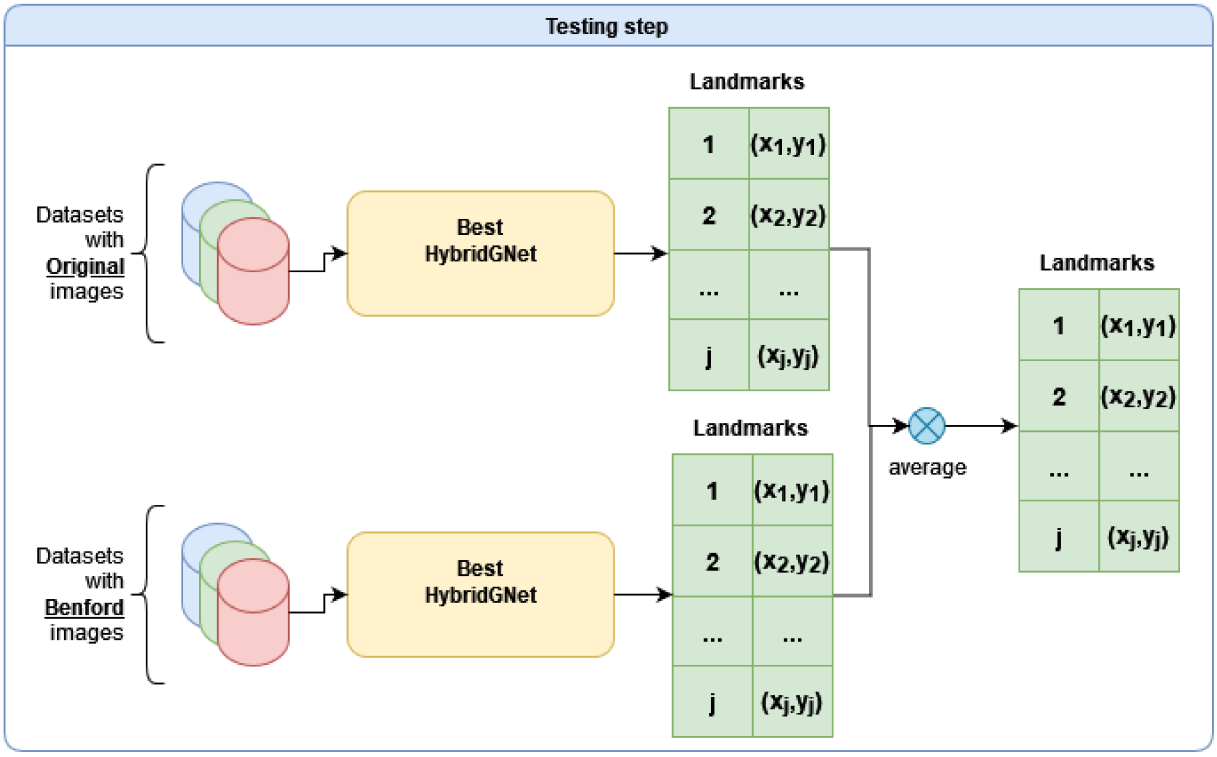


Figure 47 – Fusion process of the HybridGNet model trained with the original dataset and the model trained with the Benford dataset. The output of the models are vectors with the positions of the landmarks, the fusion and final response is the average between the vectors.

grams, dataset containing radiographs of the hand and dataset containing radiographic images of the lung, presented in Section 4.1.1, is used. The training set is composed of 130 images from the lateral cephalogram dataset, 550 images from the hand radiography dataset and 195 images from the lung radiography dataset; totaling 875 images for training. For validation, 20 lateral cephalogram images, 59 hand X-rays and 34 lung X-rays were used; a total of 113 images for validation. In the test phase, 250 lateral cephalogram images, 301 hand X-ray images and 41 lung X-ray images were processed; a total of 592 images for testing. The metrics to evaluate the model's performance is the MRE and SDR, Section 4.3.3.

When carrying out the experiments with the HybridGNet architecture, the dataset from the Japanese Society of Radiological Technology (JSRT) was used. The dataset contains 247 X-Ray images of the chest. Each image in the dataset features 166 landmarks that delimit the regions of the lungs, heart and clavicles (SHIRAISHI et al., 2000). For the experiment, the dataset was split with 70% of images for training, 10% for validation, and 20% of images for testing.

To evaluate the HybridGNet model, the Dice coefficient and Hausdorff distance (HD) metrics were used. To calculate the Dice coefficient between a source and a target image,

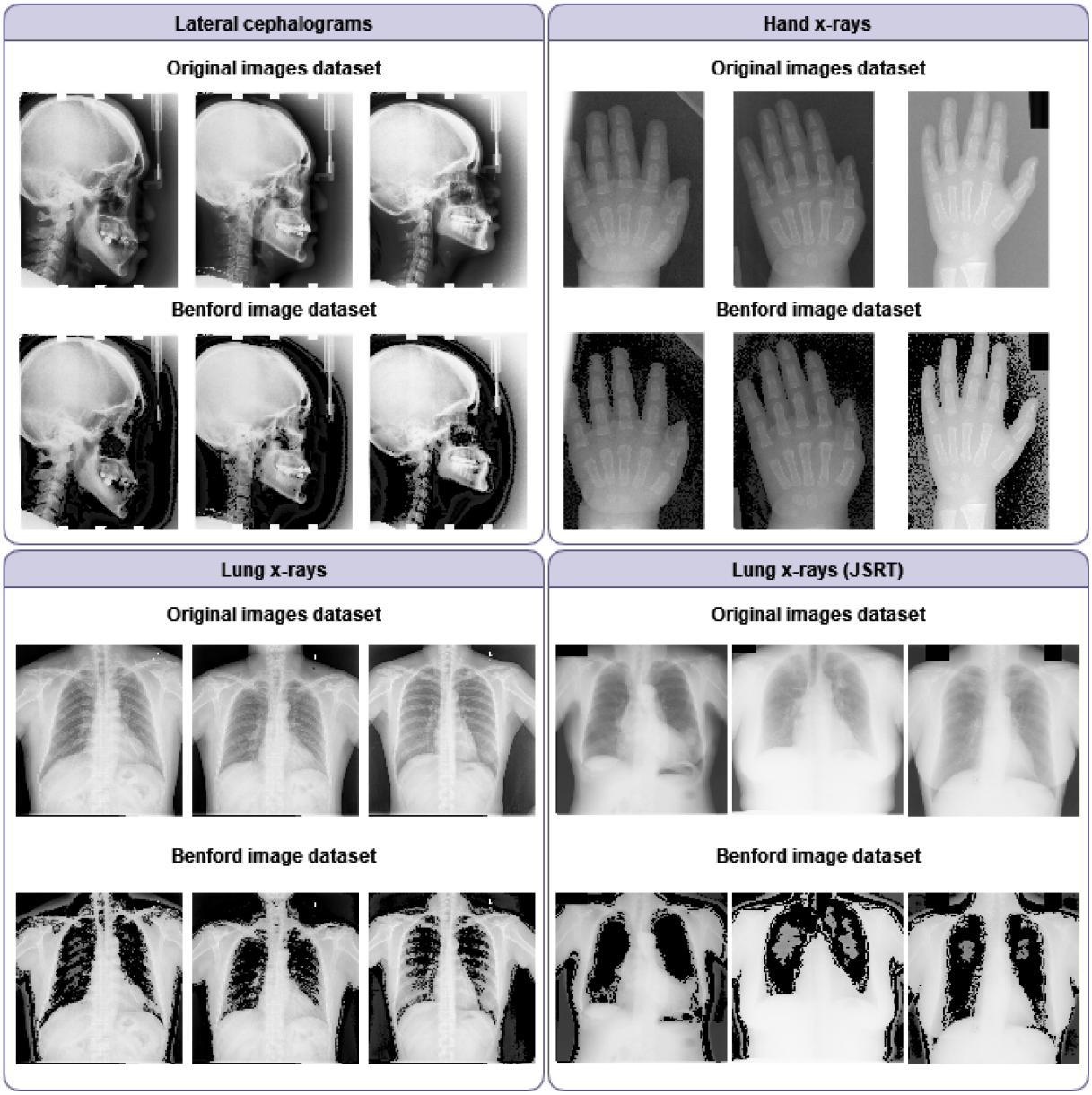


Figure 48 – Comparison of images from the original dataset with images created based on the Benford method.

the object of interest in each image is first binarized, and then the similarity between the source and target objects is measured by comparing their overlap. Hausdorff Distance is defined as the maximum distance between two objects (set of points). The distance unit is typically a value in mm (HUTTENLOCHER; KLANDERMAN; RUCKLIDGE, 1993; LI et al., 2020b; CARASS et al., 2020).

5.3.1 Experiments and results with universal model and image preprocessing

This section presents the results of the fusion of a model trained with original images and a model trained with images fitted to the Benford curve.

For each dataset the universal model (ZHU et al., 2021; ZHU et al., 2022), presented in Section 4.2.2, was trained four times for each dataset, using the random seeds 03, 42, 56, 99 and the original model parameters. Each time the model was trained for 100 epochs. Table 14 shows the average values of the results for training the model with original datasets and the model with Benford datasets.

In the testing phase, the response of each model trained with the original datasets was fused with the response of each model trained with the Benford datasets, resulting in 16 combinations. The average results of the combinations were calculated, and the final result can be seen in the last line of Table 14.

Table 14 – Universal model run results for the original datasets and for the Benford datasets

Tests	Head					Hand				Chest			
	MRE	SDR(%)				MRE	SDR(%)			MRE	SDR(%)		
	(mm)	2mm	2.5mm	3mm	4mm	(mm)	2mm	4mm	10mm	(px)	3px	6px	9px
Model with original datasets	1.48	77.95	84.24	89.55	95.09	0.76	95.45	99.01	99.80	6.25	43.09	72.87	86.18
Model with Benford datasets	1.68	72.87	80.27	86.45	93.51	0.82	95.44	99.06	99.67	5.19	41.46	77.54	88.92
Fusion of model outputs	**1.44**	**78.36**	**84.63**	**89.81**	**95.33**	**0.72**	**96.07**	**99.26**	**99.81**	**4.35**	**47.13**	**79.34**	**90.42**

The universal model was run for 100 epochs, for a mix of three datasets (lateral cephalograms, hand radiographs and lung radiographic), spent 7.52 hours, consumed 0.66 kWh and emitted 0.30 kg of $CO2eq$. This is equivalent to 2.84 km traveled by car. For model fusion, resource consumption is doubled because the model is executed twice for the same period.

5.3.2 Experiments and results with HybridGNet and image preprocessing

This section presents the results of instances of the HybridGNet model trained on two different datasets. Firstly, the model is trained and tested twice (with different random seeds) with images from the original dataset, the final result is the average of the test outputs. In the second approach, the model is trained and tested twice with images from the Benford dataset, the average of the test outputs produces the final result. In the last approach, during testing, an average fusion occurs between each landmark obtained with the model trained on the original dataset and each landmark obtained with the model trained on the Benford dataset, resulting in 4 combinations. The average result between the combinations was calculated.

In Table 15, the results of the experiment are available. Results are shown for the model being trained for 1000 epochs, 1500 epochs, and 2500 epochs. The amount of

epoch corresponds to the values presented in the HybridGNet base article. In the case of fusion, the number of epochs is doubled, as the fusion occurs between two models that were trained for the same number of epochs. The results for using the original dataset are in the first row of results. In the second line are the results related to the use of the Benford dataset. In the last line are the results referring to the fusion between the models' responses during the test. The results related to the original dataset are the best, the fusion did not produce any improvements.

Table 15 – HybridGNet model run results for the original datasets and for the Benford datasets

Tests	**Dice**			**Hausdorff (mm)**		
	1000	1500	2500	1000	1500	2500
HybridGNet with original dataset	**0.9326**	**0.9372**	**0.9387**	**14.58**	**13.79**	**13.86**
HybridGNet with Benford dataset	0.9274	0.9292	0.9314	15.88	15.58	15.20
Test	**Dice**			**Hausdorff (mm)**		
	2000	3000	5000	2000	3000	5000
HybridGNet with response fusion	0.9308	0.9354	0.9366	15.06	14.45	14.20

In Figure 49 we can see the ground truth regions, the regions obtained by the model trained with the original dataset, the regions found by the model trained with the Benford dataset and the regions delimited through fusion.

In addition to the metrics involving the quality of the markings, we consider presenting the resource consumption to train the model. To execute 1000 iterations in the HybridGNet model, 12.29 hours were spent and it is estimated that this execution uses 0.70 kWh of electricity contributing to 0.32 kg of $CO2eq$. This is equivalent to 3.01 km traveled by car. When running 1500 iterations, 18.47 hours were spent and an estimated 1.17 kWh of electricity was used, contributing to 0.54 kg of $CO2eq$. This is equivalent to 5.04 km traveled by car. In the end, after 2500 iterations, 31.04 hours were spent. It is estimated that 2.11 kWh of electricity was used, contributing to 0.98 kg of $CO2eq$. This is equivalent to 9.11 km traveled by car.

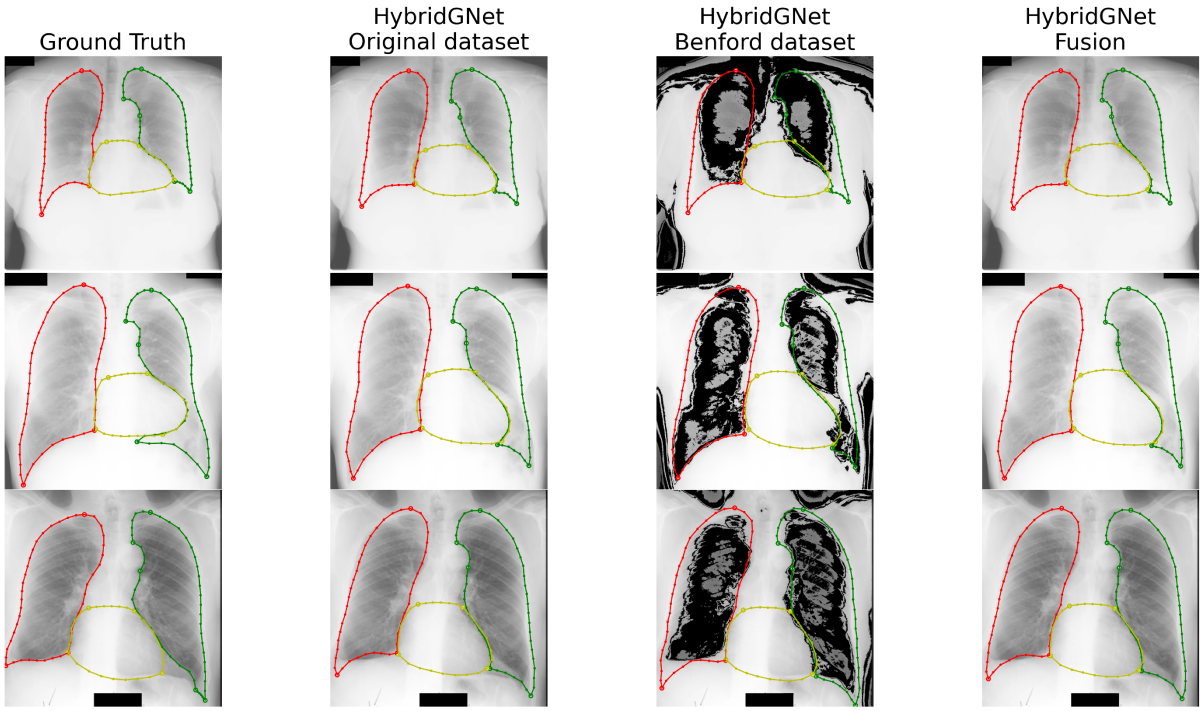


Figure 49 – From left to right: The images and markings in the first column are original. In the second column, the organs are delimited by the HybridGNet model trained with the original dataset. In the third column, the organs are delimited by the HybridGNet model trained with the Benford dataset. In the last column, the organs are delimited by fusing the responses of a HybridGNet trained with the original dataset and another HybridGNet trained with the Benford dataset.

Chapter 6

Conclusion

Technological advances have increased the use of images as an alternative for diagnosing diseases. However, human analysis of countless images can be exhaustive and prone to errors. To aid in the analysis of medical images and reduce the likelihood of errors, computer vision algorithms, particularly CNNs, have demonstrated strong performance.

These algorithms are capable of analyzing large sets of data and identifying patterns among the data that would be difficult for a human to notice. In the area of medical images, studies that seek alternatives to aid diagnosis through computer vision are important, as they can make diagnosis quick, cheap and accessible.

In this study, we analyze the performance of CNNs in classification tasks and landmark detection tasks. For the classification task, we compare different CNN ensemble architectures. We describe architectures that achieve the results shown in related work and use fewer computational resources. In classification tasks, the findings of this study highlight the significant contribution of DenseNet161 and ResNet152 to the fusion process in all experiments. Furthermore, our findings demonstrate a similar level of performance compared to the previous model, as indicated by an F1-score of 0.9100 and MCC of 0.9020. Remarkably, our approach achieves this performance using just 10GB of GPU RAM, in contrast to the previous model's requirement of 15.8GB.

In the image classification task, we concluded that it is important to evaluate different strategies, since different alternatives led to similar results. When the results are similar, it is interesting to opt for a strategy that reduces financial costs and reduces the consumption of resources such as electricity.

The Chapter 4 task is divided into two proposals. In the first proposal, we created a universal CNN model based on a network model fusion strategy. This universal model performed well in the landmark detection task. The proposed model achieved an improvement of approximately 2% in the point distance metric. Furthermore, the training is performed shorter time, which reduces energy consumption and carbon emissions by approximately 65%. For this proposal, we conclude that fusion is capable of improving performance and reducing resource consumption, however, finding a correct fusion

alternative is a difficult task.

In the Chapter 5, we seek to improve CNN performance through image preprocessing. We use original datasets and preprocessed datasets by histogram equalization and using the Benford method. We individually analyzed, for each dataset, the performance of the universal model and HybridGNet, we also evaluated the results of the models in a late fusion process. We conclude that the proposed preprocessing did not improve the model's performance. These results demonstrate the robustness of the CNNs which, for the analyzed datasets, managed to achieve good performance without the aid of the preprocessing proposal. This result does not preclude future research into other preprocessing alternatives, it only demonstrates that the histogram equalization associated with the Benford method did not bring benefits to the models analyzed.

These findings contribute to our understanding of the performance of individual models, fusion techniques, and feature utilization, paving the way for the design and implementation of more efficient and effective deep learning-based systems in image classification and landmark detection tasks.

The main challenge was accessing sufficient computing resources. Furthermore, to propose a fusion-based model there is a need for many experiments to identify the best way to combine the models. Future research should explore model compression techniques to reduce computational requirements. Furthermore, more experiments and analysis are needed to identify optimal model combinations and refine the fusion process. Evaluating fusion alternatives on different datasets is crucial for evaluating performance and resource utilization. Our study brought contributions that are presented in the next section.

6.1 Main Contributions

As the main contribution to this work, we propose a fusion-based model that achieves better accuracy than the literature and reduces energy consumption and carbon emissions by approximately 65%. The proposed model is also capable of jointly processing different datasets with different numbers of landmarks.

Furthermore, we analyze and compare different CNNs during gastrointestinal image classification. We propose, implement and compare different CNN committee architectures based on classification quality metrics and computational resource consumption.

6.2 Future works

In future work there is the possibility of applying the ensemble of classifiers to other datasets and testing other networks in the composition of the ensembles. For the landmark detection task, it is interesting to test the fusion of the universal model with the graph

spectral convolutions layers of the HybridGNet model. Furthermore, it can be assessed whether other preprocessing alternatives can help with the performance of the models.

6.3 Contributions in Bibliographic Production

- ❑ A paper was published in the BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS) (CAPES extract - A4), the title of the article is as follows: Ensemble architectures and efficient fusion techniques for Convolutional Neural Networks: an analysis on resource optimization strategies (COSTA et al., 2023).
- ❑ A paper entitled “Medical images landmarks detection by CNN fusion and energy consumption analysis” was submitted to the journal Biomedical Signal Processing and Control. The current status of the submission is under review.

Bibliography

ABDULLAH-AL-WADUD, M.; KABIR, M. H.; DEWAN, M. A. A.; CHAE, O. A dynamic histogram equalization for image contrast enhancement. **IEEE transactions on consumer electronics**, IEEE, v. 53, n. 2, p. 593–600, 2007. Disponível em: <https://doi.org/10.1109/TCE.2007.381734>.

ACEBO, E.; SBERT, M. Benford's law for natural and synthetic images. In: **Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging**. [s.n.], 2005. p. 169–176. Disponível em: <https://doi.org/10.2312/COMPAESTH/COMPAESTH05/169-176>.

ADERGHAL, K.; BENOIS-PINEAU, J.; AFDEL, K.; GWENAËLLE, C. Fuseme: Classification of smri images by fusion of deep cnns in 2d+ ε projections. In: **Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing**. [s.n.], 2017. p. 1–7. Disponível em: <https://doi.org/10.1145/3095713.3095749>.

AFIFI, M.; BROWN, M. S. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [s.n.], 2019. p. 243–252. Disponível em: <https://doi.org/10.1109/ICCV.2019.00033>.

AGRAWAL, D.; BERNSTEIN, P.; BERTINO, E.; DAVIDSON, S.; DAYAL, U.; FRANKLIN, M.; GEHRKE, J.; HAAS, L.; HALEVY, A.; HAN, J. et al. Challenges and opportunities with big data 2011-1. 2011.

AKSELA, M. Comparison of classifier selection methods for improving committee performance. In: SPRINGER. **International Workshop on Multiple Classifier Systems**. 2003. p. 84–93. Disponível em: <https://doi.org/10.1007/3-540-44938-8_9>.

AL-BANDAWI, H.; DENG, G. Classification of image distortion based on the generalized benford's law. **Multimedia Tools and Applications**, Springer, v. 78, p. 25611–25628, 2019. Disponível em: <https://doi.org/10.1007/s11042-019-7668-3>.

ALI, S.; ZHOU, F.; BRADEN, B.; BAILEY, A.; YANG, S.; CHENG, G.; ZHANG, P.; LI, X.; KAYSER, M.; SOBERANIS-MUKUL, R. D. et al. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 2748, 2020. Disponível em: <https://doi.org/10.1038/s41598-020-59413-5>.

ALWAWI, B.; ABOOD, L. Convolution neural network and histogram equalization for covid-19 diagnosis system. **Indonesian Journal of Electrical Engineering and Computer Science**, p. 420–427, 2021. Disponível em: <https://doi.org/10.11591/ijeecs.v24.i1.pp420-427>.

ANTHONY, L. F. W.; KANDING, B.; SELVAN, R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. **arXiv preprint arXiv:2007.03051**, 2020.

ARABIAN, H.; WAGNER-HARTL, V.; CHASE, J. G.; MÖLLER, K. Image pre-processing significance on regions of impact in a trained network for facial emotion recognition. **IFAC-PapersOnLine**, Elsevier, v. 54, n. 15, p. 299–303, 2021. Disponível em: <https://doi.org/10.1016/j.ifacol.2021.10.272>.

BAWDEN, D.; ROBINSON, L. Information overload: An overview. In: **Oxford Encyclopedia of Political Decision Making**. Oxford: Oxford University Press, 2020. This is a draft of a chapter that has been published by Oxford University Press in the book Oxford Encyclopedia of Political Decision Making. Disponível em: <https://doi.org/10.1093/acrefore/9780190228637.013.1360>.

BAYOUDH, K.; KNANI, R.; HAMDAOUI, F.; MTIBAA, A. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. **The Visual Computer**, p. 1–32, 2021. Disponível em: <https://doi.org/10.1007/s00371-021-02166-7>.

BENDER, E. M.; GEBRU, T.; MCMILLAN-MAJOR, A.; SHMITCHELL, S. On the dangers of stochastic parrots: Can language models be too big? In: **Proceedings of the 2021 ACM conference on fairness, accountability, and transparency**. [s.n.], 2021. p. 610–623. Disponível em: <https://doi.org/10.1145/3442188.3445922>.

BENFORD, F. The law of anomalous numbers. **Proceedings of the American Philosophical Society**, American Philosophical Society, v. 78, n. 4, p. 551–572, 1938. ISSN 0003049X. Disponível em: <http://www.jstor.org/stable/984802>.

BERGER, A.; HILL, T. P. A basic theory of Benford's Law. **Probability Surveys**, Institute of Mathematical Statistics and Bernoulli Society, v. 8, n. none, p. 1 – 126, 2011. Disponível em: <https://doi.org/10.1214/11-PS175>.

BOCHKOVSKIY, A.; WANG, C.-Y.; LIAO, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. **arXiv preprint arXiv:2004.10934**, 2020.

BORGLI, H.; THAMBAWITA, V.; SMEDSRUD, P. H.; HICKS, S.; JHA, D.; ESKELAND, S. L.; RANDEL, K. R.; POGORELOV, K.; LUX, M.; NGUYEN, D. T. D.; JOHANSEN, D.; GRIWODZ, C.; STENSLAND, H. K.; GARCIA-CEJA, E.; SCHMIDT, P. T.; HAMMER, H. L.; RIEGLER, M. A.; HALVORSEN, P.; LANGE, T. de. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. **Scientific Data**, v. 7, n. 1, p. 283, 2020. ISSN 2052-4463. Disponível em: <https://doi.org/10.1038/s41597-020-00622-y>.

BUDENNYY, S. A.; LAZAREV, V. D.; ZAKHARENKO, N. N.; KOROVIN, A. N.; PLOSSKAYA, O.; DIMITROV, D. V.; AKHRIPKIN, V.; PAVLOV, I.; OSELEDETS, I. V.; BARSOLA, I. S. et al. Eco2ai: carbon emissions tracking of

machine learning models as the first step towards sustainable ai. In: SPRINGER. **Doklady Mathematics**. 2022. v. 106, n. Suppl 1, p. S118–S128. Disponível em: <https://doi.org/10.1134/S1064562422060230>.

CANDEMIR, S.; JAEGER, S.; PALANIAPPAN, K.; MUSCO, J. P.; SINGH, R. K.; XUE, Z.; KARARGYRIS, A.; ANTANI, S.; THOMA, G.; MCDONALD, C. J. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. **IEEE transactions on medical imaging**, IEEE, v. 33, n. 2, p. 577–590, 2013. Disponível em: <https://doi.org/10.1109/TMI.2013.2290491>.

CARASS, A.; ROY, S.; GHERMAN, A.; REINHOLD, J. C.; JESSON, A.; ARBEL, T.; MAIER, O.; HANDELS, H.; GHAFOORIAN, M.; PLATEL, B. et al. Evaluating white matter lesion segmentations with refined sørensen-dice analysis. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 8242, 2020. Disponível em: <https://doi.org/10.1038/s41598-020-64803-w>.

CHEN, R.; MA, Y.; CHEN, N.; LEE, D.; WANG, W. Cephalometric landmark detection by attentive feature pyramid fusion and regression-voting. In: SPRINGER. **Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22**. 2019. p. 873–881. Disponível em: <https://doi.org/10.1007/978-3-030-32248-9_97>.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC genomics**, Springer, v. 21, p. 1–13, 2020. Disponível em: <https://doi.org/10.1186/s12864-019-6413-7>.

COSTA, C. L.; LIMA, D. A.; BARCELOS, C. A. Z.; TRAVENÇOLO, B. A. Ensemble architectures and efficient fusion techniques for convolutional neural networks: An analysis on resource optimization strategies. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. 2023. p. 107–121. Disponível em: <https://doi.org/10.1007/978-3-031-45389-2_8>.

DONG, K.; ZHOU, C.; RUAN, Y.; LI, Y. Mobilenetv2 model for image classification. In: IEEE. **2020 2nd International Conference on Information Technology and Computer Application (ITCA)**. 2020. p. 476–480. Disponível em: <https://doi.org/10.1109/ITCA52113.2020.00106>.

DURTSCHI, C.; HILLISON, W.; PACINI, C. et al. The effective use of benford's law to assist in detecting fraud in accounting data. **Journal of forensic accounting**, v. 5, n. 1, p. 17–34, 2004.

FUKUSHIMA, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. **Biological Cybernetics**, v. 36, n. 4, p. 193–202, Apr 1980. ISSN 1432-0770. Disponível em: <https://doi.org/10.1007/BF00344251>.

GAGGION, N.; MANSILLA, L.; MILONE, D. H.; FERRANTE, E. Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. In: SPRINGER. **Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24**. 2021. p. 600–610. Disponível em: <https://doi.org/10.1007/978-3-030-87193-2_57>.

GAGGION, N.; MANSILLA, L.; MOSQUERA, C.; MILONE, D. H.; FERRANTE, E. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. **IEEE Transactions on Medical Imaging**, IEEE, v. 42, n. 2, p. 546–556, 2022. Disponível em: <https://doi.org/10.1109/TMI.2022.3224660>.

GIEŁCZYK, A.; MARCINIAK, A.; TARCZEWSKA, M.; LUTOWSKI, Z. Pre-processing methods in chest x-ray image classification. **Plos one**, Public Library of Science San Francisco, CA USA, v. 17, n. 4, p. e0265949, 2022. Disponível em: <https://doi.org/10.1371/journal.pone.0265949>.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2014. p. 580–587. Disponível em: <https://doi.org/10.1109/CVPR.2014.81>.

_____. Region-based convolutional networks for accurate object detection and segmentation. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 38, n. 1, p. 142–158, 2015. Disponível em: <https://doi.org/10.1109/TPAMI.2015.2437384>.

GONZALEZ, R. C.; WOODS, R. C. **Processamento digital de imagens**. [S.l.]: Pearson Educación, 2009.

GOSHTASBY, A. A. **2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications**. [s.n.], 2005. ISBN 0471649546. Disponível em: <https://doi.org/10.1002/0471724270>.

GRAPOV, D.; FAHRMANN, J.; WANICHTHANARAK, K.; KHOOMRUNG, S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. **Omics: a journal of integrative biology**, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New, v. 22, n. 10, p. 630–636, 2018. Disponível em: <https://doi.org/10.1089/omi.2018.0097>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2016. p. 770–778. Disponível em: <https://doi.org/10.1109/CVPR.2016.90>.

HENDERSON, P.; HU, J.; ROMOFF, J.; BRUNSKILL, E.; JURAFSKY, D.; PINEAU, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. **Journal of Machine Learning Research**, v. 21, n. 248, p. 1–43, 2020. Disponível em: <http://jmlr.org/papers/v21/20-312.html>.

HICKS, S. A.; JHA, D.; THAMBAWITA, V.; HALVORSEN, P.; HAMMER, H. L.; RIEGLER, M. A. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In: SPRINGER. **Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII**. 2021. p. 263–274. Disponível em: <https://doi.org/10.1007/978-3-030-68793-9_18>.

HILL, T. P. The significant-digit phenomenon. **The American Mathematical Monthly**, Taylor & Francis, v. 102, n. 4, p. 322–327, 1995. Disponível em: <https://doi.org/10.1080/00029890.1995.11990578>.

HODAK, M.; GORKOVENKO, M.; DHOLAKIA, A. Towards power efficiency in deep learning on data center hardware. In: IEEE. **2019 IEEE International Conference on Big Data (Big Data)**. 2019. p. 1814–1820. Disponível em: <https://doi.org/10.1109/BigData47090.2019.9005632>.

HÖHN, J.; KRIEGHOFF-HENNING, E.; JUTZI, T. B.; KALLE, C. von; UTIKAL, J. S.; MEIER, F.; GELLRICH, F. F.; HOBELSBERGER, S.; HAUSCHILD, A.; SCHLAGER, J. G. et al. Combining cnn-based histologic whole slide image analysis and patient data to improve skin cancer classification. **European Journal of Cancer**, Elsevier, v. 149, p. 94–101, 2021. Disponível em: <https://doi.org/10.1016/j.ejca.2021.02.032>.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. **arXiv preprint arXiv:1704.04861**, 2017.

HUANG, G.; LIU, Z.; MAATEN, L. V. D.; WEINBERGER, K. Q. Densely connected convolutional networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2017. p. 4700–4708. Disponível em: <https://doi.org/10.1109/CVPR.2017.243>.

HUTTENLOCHER, D. P.; KLANDERMAN, G. A.; RUCKLIDGE, W. J. Comparing images using the hausdorff distance. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 15, n. 9, p. 850–863, 1993. Disponível em: <https://doi.org/10.1109/34.232073>.

IQBAL, I.; WALAYAT, K.; KAKAR, M. U.; MA, J. Automated identification of human gastrointestinal tract abnormalities based on deep convolutional neural network with endoscopic images. **Intelligent Systems with Applications**, Elsevier, v. 16, p. 200149, 2022. Disponível em: <https://doi.org/10.1016/j.iswa.2022.200149>.

JABRI, S.; DURIC, Z.; WECHSLER, H.; ROSENFELD, A. Detection and location of people in video images using adaptive fusion of color and edge information. In: IEEE. **Proceedings 15th International Conference on Pattern Recognition. ICPR-2000**. 2000. v. 4, p. 627–630. Disponível em: <https://doi.org/10.1109/ICPR.2000.902997>.

JAEGER, S.; CANDEMIR, S.; ANTANI, S.; WÁNG, Y.-X. J.; LU, P.-X.; THOMA, G. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. **Quantitative imaging in medicine and surgery**, AME Publications, v. 4, n. 6, p. 475, 2014. Disponível em: <https://doi.org/10.3978%2Fj.issn.2223-4292.2014.11.20>.

JAEGER, S.; KARARGYRIS, A.; CANDEMIR, S.; FOLIO, L.; SIEGELMAN, J.; CALLAGHAN, F.; XUE, Z.; PALANIAPPAN, K.; SINGH, R. K.; ANTANI, S. et al. Automatic tuberculosis screening using chest radiographs. **IEEE transactions on medical imaging**, IEEE, v. 33, n. 2, p. 233–245, 2013. Disponível em: <https://doi.org/10.1109/TMI.2013.2284099>.

JHA, D.; ALI, S.; TOMAR, N. K.; JOHANSEN, H. D.; JOHANSEN, D.; RITTSCHER, J.; RIEGLER, M. A.; HALVORSEN, P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. **IEEE Access**, IEEE, v. 9, p. 40496–40510, 2021. Disponível em: <https://doi.org/10.1109/ACCESS.2021.3063716>.

JHA, D.; HICKS, S. A.; EMANUELSEN, K.; JOHANSEN, H.; JOHANSEN, D.; LANGE, T. de; RIEGLER, M. A.; HALVORSEN, P. Medico multimedia task at mediaeval 2020: Automatic polyp segmentation. **arXiv preprint arXiv:2012.15244**, 2020.

JHA, D.; SMEDSRUD, P. H.; JOHANSEN, D.; LANGE, T. de; JOHANSEN, H. D.; HALVORSEN, P.; RIEGLER, M. A. A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. **IEEE journal of biomedical and health informatics**, IEEE, v. 25, n. 6, p. 2029–2040, 2021. Disponível em: <https://doi.org/10.1109/JBHI.2021.3049304>.

JHA, D.; SMEDSRUD, P. H.; RIEGLER, M. A.; HALVORSEN, P.; LANGE, T. de; JOHANSEN, D.; JOHANSEN, H. D. Kvasir-seg: A segmented polyp dataset. In: SPRINGER. **MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26**. 2020. p. 451–462. Disponível em: <https://doi.org/10.1007/978-3-030-37734-2_37>.

JO, T.; NHO, K.; SAYKIN, A. J. Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. **Frontiers in aging neuroscience**, Frontiers, v. 11, p. 220, 2019. Disponível em: <https://doi.org/10.3389/fnagi.2019.00220>.

JOLION, J.-M. Images and benford's law. **Journal of Mathematical Imaging and Vision**, Springer, v. 14, p. 73–81, 2001. Disponível em: <https://doi.org/10.1023/A:1008363415314>.

KESSLER, M. M. Bibliographic coupling between scientific papers. **American documentation**, Wiley Online Library, v. 14, n. 1, p. 10–25, 1963. Disponível em: <https://doi.org/10.1002/asi.5090140103>.

KIM, B.-K.; LEE, H.; ROH, J.; LEE, S.-Y. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In: **Proceedings of the 2015 ACM on International Conference on Multimodal Interaction**. [s.n.], 2015. p. 427–434. Disponível em: <https://doi.org/10.1145/2818346.2830590>.

KRAWCZYK, B.; SCHAEFER, G. A hybrid classifier committee for analysing asymmetry features in breast thermograms. **Applied Soft Computing**, Elsevier, v. 20, p. 112–118, 2014. Disponível em: <https://doi.org/10.1016/j.asoc.2013.11.011>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, p. 1097–1105, 2012. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.

LACOSTE, A.; LUCCIONI, A.; SCHMIDT, V.; DANDRES, T. Quantifying the carbon emissions of machine learning. **arXiv preprint arXiv:1910.09700**, 2019.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. Disponível em: <https://doi.org/10.1038/nature14539>.

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. **Neural computation**, MIT Press, v. 1, n. 4, p. 541–551, 1989. Disponível em: <https://doi.org/10.1162/neco.1989.1.4.541>.

LEE, J.-H.; YU, H.-J.; KIM, M.-j.; KIM, J.-W.; CHOI, J. Automated cephalometric landmark detection with confidence regions using bayesian convolutional neural networks. **BMC oral health**, Springer, v. 20, p. 1–10, 2020. Disponível em: <https://doi.org/10.1186/s12903-020-01256-7>.

LI, W.; ZHUANG, J.; WANG, R.; ZHANG, J.; ZHENG, W.-S. Fusing metadata and dermoscopy images for skin disease diagnosis. In: IEEE. **2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)**. 2020. p. 1996–2000. Disponível em: <https://doi.org/10.1109/ISBI45749.2020.9098645>.

LI, X.; SUN, X.; MENG, Y.; LIANG, J.; WU, F.; LI, J. Dice loss for data-imbalanced NLP tasks. In: JURAFSKY, D.; CHAI, J.; SCHLUTER, N.; TETREAULT, J. (Ed.). **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 465–476. Disponível em: <https://aclanthology.org/2020.acl-main.45>.

LIN, S.-Y.; CHANG, H.-Y. Tooth numbering and condition recognition on dental panoramic radiograph images using cnns. **IEEE Access**, IEEE, v. 9, p. 166008–166026, 2021. Disponível em: <https://doi.org/10.1109/ACCESS.2021.3136026>.

LINDNER, C.; WANG, C.-W.; HUANG, C.-T.; LI, C.-H.; CHANG, S.-W.; COOTES, T. F. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. **Scientific reports**, Nature Publishing Group UK London, v. 6, n. 1, p. 33581, 2016. Disponível em: <https://doi.org/10.1038/srep33581>.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. V. D.; GINNEKEN, B. V.; SÁNCHEZ, C. I. A survey on deep learning in medical image analysis. **Medical image analysis**, Elsevier, v. 42, p. 60–88, 2017. Disponível em: <https://doi.org/10.1016/j.media.2017.07.005>.

MAKRUSHIN, A.; KRAETZER, C.; NEUBERT, T.; DITTMANN, J. Generalized benford's law for blind detection of morphed face images. In: **Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security**. [s.n.], 2018. p. 49–54. Disponível em: <https://doi.org/10.1145/3206004.3206018>.

MASLEJ, N.; FATTORINI, L.; PERRAULT, R.; PARLI, V.; REUEL, A.; BRYNJOLFSSON, E.; ETCHEMENDY, J.; LIGETT, K.; LYONS, T.; MANYIKA, J.; NIEBLES, J.; SHOHAM, Y.; WALD, R.; CLARK, J. **Artificial Intelligence Index Report 2024**. Institute for Human-Centered AI, Stanford University, 2024. Disponível em: <https://coilink.org/20.500.12592/h70s46h>.

MESQUITA, G. de Q. T. B.; VIEIRA, W. A.; VIDIGAL, M. T. C.; TRAVENÇOLO, B. A. N.; BEAINI, T. L.; SPIN-NETO, R.; PARANHOS, L. R.; JÚNIOR, R. B. de B.

Artificial intelligence for detecting cephalometric landmarks: A systematic review and meta-analysis. **Journal of Digital Imaging**, Springer, v. 36, n. 3, p. 1158–1179, 2023. Disponível em: <https://doi.org/10.1007/s10278-022-00766-w>.

MODERSITZKI, J. **FAIR: flexible algorithms for image registration**. Siam, 2009. v. 6. Disponível em: <https://doi.org/10.1137/1.9780898718843>.

MURPHY, K.; TORRALBA, A.; EATON, D.; FREEMAN, W. Object detection and localization using local and global features. **Toward category-level object recognition**, Springer, p. 382–400, 2006. Disponível em: <https://doi.org/10.1007/11957959_20>.

NAESS, E.; THAMBAWITA, V.; HICKS, S. A.; RIEGLER, M. A.; HALVORSEN, P. Pyramidal segmentation of medical images using adversarial training. In: **Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval**. [s.n.], 2021. p. 33–38. Disponível em: <https://doi.org/10.1145/3463944.3469100>.

OSAREH, F. Bibliometrics, citation analysis and co-citation analysis: A review of literature i. Walter de Gruyter, Berlin/New York Berlin, New York, 1996. Disponível em: <https://doi.org/10.1515/libr.1996.46.3.149>.

ÖZTÜRK, Ş.; AKDEMIR, B. Effects of histopathological image pre-processing on convolutional neural networks. **Procedia computer science**, Elsevier, v. 132, p. 396–403, 2018. Disponível em: <https://doi.org/10.1016/j.procs.2018.05.166>.

PAPANDREOU, G.; ZHU, T.; KANAZAWA, N.; TOSHEV, A.; TOMPSON, J.; BREGLER, C.; MURPHY, K. Towards accurate multi-person pose estimation in the wild. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2017. p. 4903–4911. Disponível em: <https://doi.org/10.1109/CVPR.2017.395>.

PARAGIOS, N.; TZIRITAS, G. Adaptive detection and localization of moving objects in image sequences. **Signal Processing: Image Communication**, Elsevier, v. 14, n. 4, p. 277–296, 1999. Disponível em: <https://doi.org/10.1016/S0923-5965(98)00011-3>.

PARNAK, A.; DAMAVANDI, Y. B.; KAZEMITABAR, S. A novel image splicing detection algorithm based on generalized and traditional benford's law. **International Journal of Engineering**, Materials and Energy Research Center, v. 35, n. 4, p. 626–634, 2022. Disponível em: <https://doi.org/10.5829/IJE.2022.35.04A.02>.

PAYER, C.; ŠTERN, D.; BISCHOF, H.; URSCHLER, M. Regressing heatmaps for multiple landmark localization using cnns. In: SPRINGER. **International conference on medical image computing and computer-assisted intervention**. 2016. p. 230–238. Disponível em: <https://doi.org/10.1007/978-3-319-46723-8_27>.

_____. Integrating spatial configuration into heatmap regression based cnns for landmark localization. **Medical image analysis**, Elsevier, v. 54, p. 207–219, 2019. Disponível em: <https://doi.org/10.1016/j.media.2019.03.007>.

PIZER, S. M.; AMBURN, E. P.; AUSTIN, J. D.; CROMARTIE, R.; GESELOWITZ, A.; GREER, T.; ROMENY, B. ter H.; ZIMMERMAN, J. B.; ZUIDERVELD, K. Adaptive histogram equalization and its variations. **Computer vision, graphics,**

and image processing, Elsevier, v. 39, n. 3, p. 355–368, 1987. Disponível em: <https://doi.org/10.1016/S0734-189X(87)80186-X>.

RADEVSKI, V.; BENNANI, Y. Reliability control in committee classifier environment. In: IEEE. **Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium**. 2000. v. 3, p. 561–565. Disponível em: <https://doi.org/10.1109/IJCNN.2000.861369>.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You only look once: Unified, real-time object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2016. p. 779–788. Disponível em: <https://doi.org/10.1109/CVPR.2016.91>.

REDMON, J.; FARHADI, A. Yolo9000: better, faster, stronger. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2017. p. 7263–7271. Disponível em: <https://doi.org/10.1109/CVPR.2017.690>.

_____. Yolov3: An incremental improvement. **arXiv preprint arXiv:1804.02767**, 2018.

RUNDO, L.; MILITELLO, C.; VITABILE, S.; RUSSO, G.; SALA, E.; GILARDI, M. C. A survey on nature-inspired medical image analysis: a step further in biomedical data integration. **Fundamenta Informaticae**, IOS Press, v. 171, n. 1-4, p. 345–365, 2020. Disponível em: <https://doi.org/10.3233/FI-2020-1887>.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M. et al. Imagenet large scale visual recognition challenge. **International journal of computer vision**, Springer, v. 115, p. 211–252, 2015. Disponível em: <https://doi.org/10.1007/s11263-015-0816-y>.

SAMBRIDGE, M.; TKALČIĆ, H.; JACKSON, A. Benford's law in the natural sciences. **Geophysical research letters**, Wiley Online Library, v. 37, n. 22, 2010. Disponível em: <https://doi.org/10.1029/2010GL044830>.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [s.n.], 2018. p. 4510–4520. Disponível em: <https://doi.org/10.1109/CVPR.2018.00474>.

SANTOS, D. F. dos; FARIA, P. R. de; TRAVENÇOLO, B. A.; NASCIMENTO, M. Z. do. Automated detection of tumor regions from oral histological whole slide images using fully convolutional neural networks. **Biomedical Signal Processing and Control**, Elsevier BV, v. 69, p. 102921, aug 2021. Disponível em: <https://doi.org/10.1016/j.bspc.2021.102921>.

SARKAR, D.; BALI, R.; SHARMA, T. **Practical Machine Learning with Python**. [s.n.], 2018. ISBN 978-1-4842-3206-4. Disponível em: <https://doi.org/10.1007/978-1-4842-3207-1>.

SELVAN, R.; BHAGWAT, N.; ANTHONY, L. F. W.; KANDING, B.; DAM, E. B. Carbon footprint of selecting and training deep learning models for medical image analysis. In: SPRINGER. **International Conference on Medical Image**

Computing and Computer-Assisted Intervention. 2022. p. 506–516. Disponível em: <https://doi.org/10.1007/978-3-031-16443-9_49>.

SHIN, M.; KIM, M.; KWON, D.-S. Baseline cnn structure analysis for facial expression recognition. In: IEEE. **2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)**. 2016. p. 724–729. Disponível em: <https://doi.org/10.1109/ROMAN.2016.7745199>.

SHIRAISHI, J.; KATSURAGAWA, S.; IKEZOE, J.; MATSUMOTO, T.; KOBAYASHI, T.; KOMATSU, K.-i.; MATSUI, M.; FUJITA, H.; KODERA, Y.; DOI, K. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. **American journal of roentgenology**, Am Roentgen Ray Soc, v. 174, n. 1, p. 71–74, 2000. Disponível em: <https://doi.org/10.2214/ajr.174.1.1740071>.

SONKA, M.; HLAVAC, V.; BOYLE, R.; SONKA, M.; HLAVAC, V.; BOYLE, R. Image pre-processing. **Image processing, analysis and machine vision**, Springer, p. 56–111, 1993. Disponível em: <https://doi.org/10.1007/978-1-4899-3216-7_4>.

STOCKMAN, G.; SHAPIRO, L. G. **Computer vision**. [S.l.]: Prentice Hall PTR, 2001. 1–608 p. ISBN 0130307963.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. Energy and policy considerations for deep learning in NLP. In: KORHONEN, A.; TRAUM, D.; MÀRQUEZ, L. (Ed.). **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 3645–3650. Disponível em: <https://aclanthology.org/P19-1355>.

SWENSSON, R. G. Unified measurement of observer performance in detecting and localizing target objects on images. **Medical physics**, Wiley Online Library, v. 23, n. 10, p. 1709–1725, 1996. Disponível em: <https://doi.org/10.1118/1.597758>.

SZE, V.; CHEN, Y.-H.; YANG, T.-J.; EMER, J. S. Efficient processing of deep neural networks: A tutorial and survey. **Proceedings of the IEEE**, IEEE, v. 105, n. 12, p. 2295–2329, 2017. Disponível em: <https://doi.org/10.1109/JPROC.2017.2761740>.

_____. Efficient processing of deep neural networks. **Synthesis Lectures on Computer Architecture**, Morgan & Claypool Publishers, v. 15, n. 2, p. 1–341, 2020. Disponível em: <https://doi.org/10.2200/S01004ED1V01Y202004CAC050>.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: **Thirty-first AAAI conference on artificial intelligence**. [s.n.], 2017. Disponível em: <https://doi.org/10.1609/aaai.v31i1.11231>.

TABIK, S.; PERALTA, D.; HERRERA-POYATOS, A.; TRIGUERO, F. H. et al. A snapshot of image pre-processing for convolutional neural networks: case study of mnist. Atlantis Press, 2017. Disponível em: <https://doi.org/10.2991/ijcis.2017.10.1.38>.

TAKAHASHI, K.; YAMAMOTO, K.; KUCHIBA, A.; KOYAMA, T. Confidence interval for micro-averaged f 1 and macro-averaged f 1 scores. **Applied Intelligence**, Springer, v. 52, n. 5, p. 4961–4972, 2022. Disponível em: <https://doi.org/10.1007/s10489-021-02635-5>.

TAMMINA, S. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. **International Journal of Scientific and Research Publications (IJSRP)**, International Journal of Scientific and Research Publications (IJSRP), v. 9, n. 10, p. 143–150, 2019. Disponível em: <https://doi.org/10.29322/IJSRP.9.10.2019.p9420>.

TAN, M.; LE, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). **Proceedings of the 36th International Conference on Machine Learning**. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 6105–6114. Disponível em: <https://proceedings.mlr.press/v97/tan19a.html>.

THAMBAWITA, V.; JHA, D.; HAMMER, H. L.; JOHANSEN, H. D.; JOHANSEN, D.; HALVORSEN, P.; RIEGLER, M. A. An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification. **ACM Transactions on Computing for Healthcare**, ACM New York, NY, USA, v. 1, n. 3, p. 1–29, 2020. Disponível em: <https://doi.org/10.1145/3386295>.

VADDI, R.; MANOHARAN, P. Hyperspectral image classification using cnn with spectral and spatial features integration. **Infrared Physics & Technology**, Elsevier, v. 107, p. 103296, 2020. Disponível em: <https://doi.org/10.1016/j.infrared.2020.103296>.

VOULODIMOS, A.; DOULAMIS, N.; DOULAMIS, A.; PROTOPAPADAKIS, E. Deep learning for computer vision: A brief review. **Computational intelligence and neuroscience**, Wiley Online Library, v. 2018, n. 1, p. 7068349, 2018. Disponível em: <https://doi.org/10.1155/2018/7068349>.

WANG, C.-Y.; BOCHKOVSKIY, A.; LIAO, H.-Y. M. Scaled-YOLOv4: Scaling cross stage partial network. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [s.n.], 2021. p. 13029–13038. Disponível em: <https://doi.org/10.1109/CVPR46437.2021.01283>.

_____. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [s.n.], 2023. p. 7464–7475. Disponível em: <https://doi.org/10.1109/CVPR52729.2023.00721>.

WANG, C.-Y.; LIAO, H.-Y. M.; YEH, I.-H. Designing network design strategies through gradient path analysis. **arXiv preprint arXiv:2211.04800**, 2022.

WELLS, K.; CHIVERTON, J.; PARTRIDGE, M.; BARRY, M.; KADHEM, H.; OTT, B. Quantifying the partial volume effect in pet using benford's law. **IEEE Transactions on Nuclear science**, IEEE, v. 54, n. 5, p. 1616–1625, 2007. Disponível em: <https://doi.org/10.1109/TNS.2007.903182>.

WOODS, J. W. **Multidimensional Signal, Image, and Video Processing and Coding**. [s.n.], 2011. ISBN 9780123814203. Disponível em: <https://doi.org/10.1016/C2009-0-62200-5>.

ZENG, M.; YAN, Z.; LIU, S.; ZHOU, Y.; QIU, L. Cascaded convolutional networks for automatic cephalometric landmark detection. **Medical Image Analysis**, Elsevier, v. 68, p. 101904, 2021. Disponível em: <https://doi.org/10.1016/j.media.2020.101904>.

ZHU, H.; YAO, Q.; XIAO, L.; ZHOU, S. K. You only learn once: Universal anatomical landmark detection. In: SPRINGER. **Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24**. 2021. p. 85–95. Disponível em: <https://doi.org/10.1007/978-3-030-87240-3_9>.

_____. Learning to localize cross-anatomy landmarks in x-ray images with a universal model. **BME Frontiers**, AAAS, v. 2022, 2022. Disponível em: <https://doi.org/10.34133/2022/9765095>.