Deep learning and data warehousing techniques applied to real data in the medical domain

Daniel Mário de Lima

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Daniel Mário de Lima

Deep learning and data warehousing techniques applied to real data in the medical domain

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. José Fernando Rodrigues Junior

USP – São Carlos
March 2023

Daniel Mário de Lima

Técnicas de deep learning e data warehousing aplicadas a dados reais do domínio médico

USP – São Carlos
Março de 2023

In memory of those lost to disease, poverty and ignorance, when illusions subdue judgment.

ACKNOWLEDGEMENTS

"O amor tudo pode,
para ele não há obstáculos de raça, de fortuna, de condição;
ele vence, com ou sem pretor, zomba da Igreja e da Fortuna,
e o estado amoroso é a maior delícia da nossa existência,
que se deve procurar gozá-lo e sofrê-lo,
seja como for.
O martírio até dá-lhe mais requinte..."
(Lima Barreto)

RESUMO

LIMA, D. M. **Técnicas de deep learning e data warehousing aplicadas a dados reais do domínio médico**. 2023. 80 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Este estudo visa ampliar o aproveitamento dos dados médicos e da capacidade de diagnóstico automatizado através da integração e homogeneização das diversas fontes de dados proveniente do Sistema de Informações de Saúde SI3 do Instituto do Coração (InCor/HC.FMUSP), e a investigação de modelos do estado-da-arte de aprendizado de máquina conhecidos por Deep Learning, avaliando o potencial do Deep Learning de auxílio ao diagnóstico computadorizado. Como resultados, foi preparado uma base de dados para pesquisa clínica em formato OMOP-CDM, denominado InCor-CDM. No segundo artigo obteve-se até 91% de acurácia na classificação de lesões cutâneas usando uma rede neural convolucional profunda sobre a base de dados de imagens dermatoscópicas ISIC. E no terceiro artigo melhorou-se, em média, a segmentação de imagens de ressonância magnética cardíaca em 1,7% na métrica Dice e 2,5x em velocidade de treinamento de uma rede neural convolucional U-Net usando um algoritmo de localização. Estes resultados demonstram etapas de preparação de dados; aprendizagem profunda aplicada a conceitos médicos de alto nível — multi-classificação voltada a diagnóstico; e aprendizagem profunda aplicada em dados de baixo nível — segmentação de imagens de RM Cardíaca.

Palavras-chave: Aprendizagem profunda, Armazém de dados, Pesquisa Clínica, Dermatoscopia, RM Cardíaca.

ABSTRACT

LIMA, D. M. **Deep learning and data warehousing techniques applied to real data in the medical domain**. 2023. 80 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

This study aims to increase the use of medical data and the ability to automated diagnosis through the integration and homogenization of the databases from the SI3 Health Information System of the Heart Institute (InCor / HC.FMUSP), and investigate the application of state-of-the-art machine learning models known as Deep Learning, assessing the potential of Deep Learning to computerized diagnosis. As results, a database was prepared for clinical research in the OMOP-CDM format, called InCor-CDM. In the second study we obtained up to 91% overall accuracy in the classification of cutaneous lesions using a deep convolutional neural network on the ISIC database of dermatoscopic images. In the third paper we improved the segmentation of heart magnetic resonance images, on average, by 1.7% in the Dice metric and 2.5x in the training speed of a U-Net convolutional neural network using a localization algorithm. These results demonstrate steps of data preparation; deep learning applied to high-level medical concepts — multi-classification for diagnosis; and deep learning applied to low-level image data – Cardiac MRI image segmentation.

Keywords: Deep learning, Data warehouse, Clinical research, Dermatoscopy, Cardiac MRI.

CONTENTS

INTRODUCTION

1.1 Context and Problem

Deep Learning (DL) is a branch of Artificial Intelligence (AI), and studies large neural network models that learn complex tasks from data. LeCun, Bengio and Hinton (2015) note that DL's key feature is discovering features automatically, which was believed to be infeasible until the end of the 90s — the models would be computationally expensive and have ill-defined objective functions that would get stuck to local minima.

In the last decade however, Deep Neural Networks (DNNs) used in DL have reached levels of unprecedented performance due to two factors: the explosion in data production and the popularization of high-density parallel hardware. These phenomena are known as *Big Data* as reported in the database community by Agrawal *et al.* (2008), and *General-Purpose computing on Graphics Processing Units* (GPGPU) as in the survey of Wu and Liu (2008).

Big Data is characterized by many **V**s of data processing: volume, variety, velocity, value and veracity which are usually recognized in the database community as Oweis *et al.* (2015) also noted. On the other hand, Wu and Liu (2008) showed how GPGPU was a conceptual paradigm shift in Computer Graphics that allowed any parallel computation on n-dimensional arrays, replacing the previous generation of shading languages that were designed for graphics only.

Those achievements turned out to revive DL research, when Krizhevsky, Sutskever and Hinton (2012) managed to beat the state-of-the-art using a deep Convolutional Neural Network (CNN) for classification of 1.2 million images in a thousand classes — the ImageNet Large-Scale Visual Recognition Challenge (LSVRC) 2010. Most editions of ImageNet LSVRC since then had DNNs as the top performing models, and several other challenges were proposed as well.

A natural consequence of DL's success was the experimentation of its techniques in other sciences, such as Biology as shown by Tang *et al.* (2019), Medicine as in the works of

Esteva *et al.* (2017) and Wang, Casalino and Khullar (2019), and Physics as in the paper of Raissi (2018) — which hints DL applicability to dynamic systems in general. Geyer and Carle (2018) shows an application of DL to the classic routing problem, using a graph-based DNN model.

Those previous results from literature encouraged DL usage in more areas and databases, as data production and storage only grows. One interesting area for DL is healthcare, largely due to the availability of real medical data in Electronic Health Records (EHR) systems, along the pioneering work in medical imaging and informatics since the 70s. Henry *et al.* (2016) cites that 84% of hospitals in the U.S.A. had an EHR by 2015.

EHRs are part of a broad group of Health Information Systems, which also contains Hospital Information Systems (HIS), Electronic Medical Records (EMR) and Electronic Patient Records (EPR). Garets and Davis (2005) distinguishes EMRs and EHRs from a semantic standpoint: EMRs are computerized legal clinical records of the patients, while the EHR is a subset used for data exchange with other partners, e.g. government and insurance firms.

HIS are information systems with a broader scope, as Gardner, Pryor and Warner (1999) shows. It covers not only medical data, but also supports other hospital's activities such as accounting, finance, scheduling and administrative tasks. Furuie *et al.* (2007) defines the EPR as the integration of a HIS and a Picture Archiving and Communication System (PACS), offering the complete patient clinical history and all medical images in one system.

Using DL with real data requires a preliminary data preparation with tools such as Data Warehouses (DWs), the second topic of this thesis. Data warehousing is defined by Inmon (2005) as a database architecture for integration of data sources, and that stores granular data for the On-Line Analytical Processing (OLAP) aggregation layer. With the Big Data explosion, Agrawal *et al.* (2008) remarks that it became increasingly difficult to integrate and curate DWs, due to the inherent complexity of manipulating several database systems and models.

One approch for DW preparation in the health sciences is the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), described by Hripcsak *et al.* (2015). The OMOP group was later renamed to Observational Health Data Sciences and Informatics (OHDSI), and the OMOP CDM is their collaborative database model that captures all medical information around the patient. OMOP CDM also includes controlled medical vocabularies, so that partners can translate their data and participate in large multi-centric clinical studies.

Given these technologies and methodological advances, it seems possible to apply DL to medical data with a reasonable effort. The main challenges are first the consolidation of the data in the medical records, and second the definition of DL models that answer medical questions in a logical sequence, using the appropriate domain-specific language.

1.1.1   Health Information Systems

Health Information Systems, especially EPR systems, have become a widespread technology in hospitals, supporting the technical, medical and nursing staff in capturing data from health, storage, communication, analysis, diagnosis and in the organization of the flow of clinical procedures work. Despite initial resistance to the adoption of this technology, research by Farias *et al.* (2011), Goldberg *et al.* (2012), Henry *et al.* (2016) on the adoption of EPRs has demonstrated a perceived improvement in management workload, operational costs and information control, facilitating the standardization of clinical routines and strategic planning.

In 2017, Abrahão, Nobre and Gutierrez (2017) demonstrated the use of data from the EPR implemented at the Instituto do Coração (InCor/HC.FMUSP) for clinical studies based on cohort selection. That study is the reference InCor data source for clinical research, named Pauá, and it was created by defining materialized views over a database previously initialized in an extract-transform-load (ETL) process and an anonymization strategy that deleted private identifiers then scrambled protected health information (PHI).

The main limitation of the Pauá database was the lack of linkage to raw medical data (such as signals and images), as the database was created to manipulate data in the clinical information level — that is, focused on categorical diagnostics and one-dimensional quantities. So, the first step to extend Pauá effort is to create a data warehouse that supports all Pauá functions and also support the preparation of raw data for the DL models.

1.1.2   Biomedical Informatics

Biomedical Informatics is traditionally concerned with a wide range of Informatics applications in the Healthcare domain. Macedo *et al.* (2021) describe an example of a telemedicine service that was fully supported by Informatics, providing multidisciplinary training and remote followup of critical patients. From the Informatics side, that line of work involves quite common workflows, such as setting up virtual networks, developing systems and dashboards, training users and creating statistical reports for scientific publications.

Another line of Biomedical Informatics work is the development of computer methods for healthcare applications. The workflow includes a complete data analysis framework, but adapted to the medical domain. Lima *et al.* (2021) describe the analysis of COVID-19 data from medical reports considering patient conditions and symptoms, resulting in a Machine Learning (ML) model to predict the probability that a person was infected with the SARS-CoV2 virus.

That study compared the symptoms of the COVID-19 disease to previous reports of other respiratory diseases, namely the swine and bird flu outbreaks caused by subtypes of Influenza virus. The ML model was embedded in a web application that allowed healthcare workers and managers to observe the infections rates inside a hospital, and it was also adopted as a first line of defense to triage persons before entering the facilities.
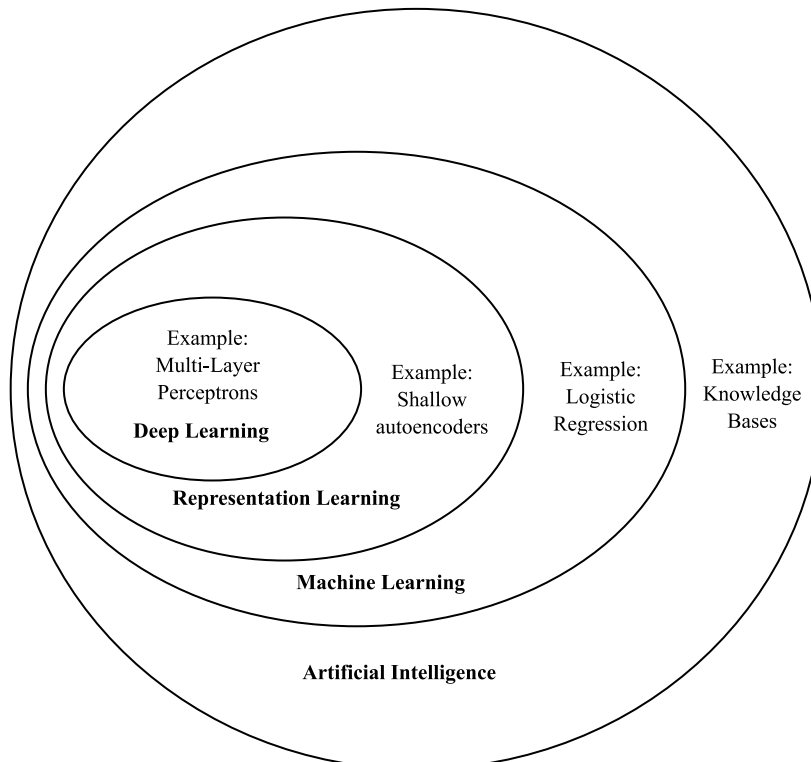
Figure 1 – Artificial Intelligence and some of its subfields, displaying the relationship between DL, RL, ML and AI. Adapted from Goodfellow, Bengio and Courville (2016).

However, even though those applications are interesting to the medical community, they use small datasets and simple statistical models, which makes it difficult to extrapolate results to general scenarios such as population health. These two points motivate experiments in larger databases and with more detailed DL models, testing if DL has the ability to improve diagnostic results or propose new healthcare applications.

1.1.3   Artificial Intelligence

Goodfellow, Bengio and Courville (2016) defines DL as a subfield of AI that is concerned with the study of programs that learn knowledge from experience using computational models based on large and complex graphs, where concepts are learned in layered subgraphs. Figure 1 illustrates the relationships between DL, AI and intermediate subfields.

From that perspective, DL composes more abstract concept representations from less abstract representations, where *representation* is a mathematical object that succintly describes another set of objects (such as images), and whose algorithms are studied in Representation Learning (RL). As an RL example, one can think of the classic Principal Component Analysis (PCA), which learns a low-dimensional representation aligned to the axes of highest variance.

Both DL and RL are inside Machine Learning (ML), which studies algorithms that can learn knowledge from data, in the form of mathematical and statistical models. Linear regressions are a classic example where the knowledge lies in the model parameters, which

are discovered by minimizing an error function to the input data. On the other hand, AI encompasses all forms of representing knowledge and intelligence in the computer, such as knowledge bases, rule-based systems and logic programming. AI history can be traced back to the first attempts of understand intelligence, for example with the development of neural networks.

Artificial Neural Networks (ANNs) are a set of ML models inspired by the functioning of the neuron, a type of cell that makes up the nervous system of animals and is responsible for conducting electrical currents between different parts of the body. Gotch and Horsley (1892) studied the anatomy and functioning of these cells around the 18th century. After the turn of the century, Rashevsky (1935) reported growing interest in the mathematical analysis of biophysical systems. After a few years, McCulloch and Pitts (1943) developed a mathematical network of neuron functions that is able to compute logical expressions, and thus computationally equivalent to the Turing (1936) machine. Then, Rosenblatt (1958) proposed another neural model, the Perceptron, which is widely recognized as seminal work for ANNs.

In the following decades, ANNs gained the ability to be trained from data and received a wide arrangement of architectures and applications. LeCun *et al.* (1989) demonstrated both the back-propagation training and architectural changes that were essential to the next wave of deep CNNs such as the AlexNet from Krizhevsky, Sutskever and Hinton (2012). Since then, DL had solid success in learning from large image datasets, which motivates its application to complex real-world scenarios such as medical images alongside text and raw signals.

1.2 Motivation

The application of DL to digital images of the skin can aid in the diagnosis. Initialization and prioritization of patient transfers in locations with difficult access to dermatologists and thus reduce melanoma mortality. This is a disease readily curable if recognized and treated in its early stages. Digital images of skin lesions have the potential to educate health professionals and the public in the identification of melanoma, as well as directly assisting in the diagnosis of melanoma through teledermatology, clinical decision support and automated diagnosis.

Other medical areas such as cardiology could benefit as well. In Cardiology, the types of images and exams are completely different from the dermatology scenario, but they require the same sort of computer routines for image processing and image analysis. Also, DL methods are general computing models that allow other forms of data to be processed and integrated together, such as body temperature signals and diagnostic codes.

The general objective of this work is to expand the analysis of medical data for the automatic diagnosis. This analysis is carried out in several stages: initial data acquisition, data storage, pre-processing, data confirmation (cleaning, normalization, selection, transformation, etc.), definition of diagnostic hypotheses by specialists, and automated reporting. These steps

are adapted from Fayyad, Piatetsky-Shapiro and Smyth (1996) to a medical framework. In the medical analysis framework, the initial data acquisition is guided by clinical estimates and employ physician anamnesis, medical imaging, laboratory tests, and other procedures.

The data is then stored in structured databases (e.g. HIS, EHR, EMR or EPRs), semi-structured (e.g. diagnostic standards such as the ICD by the World Health Organization (2016)) or unstructured (e.g. free text report written by the clinician during a patient visit, computerized X-ray slides, and electric signals captured by electrocardiography). The diagnostic hypothesis is defined with patient cohort selection; grouping and visualization of textual, numerical or complex features; survival curves; prognostic models; among other Biostatistics methods.

The algorithms used by the diagnostic methods are studied in several disciplines in Computer Science, presenting interdisciplinarity regarding the objects under analysis (medical databases) and multidisciplinarity due to the common foundation in the theories of Computing and other exact sciences. This context motivates the study of DL both in Medicine and in Computer Science, as it can help health professionals to save lives and it also provides real-world data for algorithmic research in Computer Science, Mathematics and Statistics.

1.3   Questions and Hypotheses

Concerning the complexity of medical databases and the potential applicability of DL to both medical images and clinical data, this thesis is motivated by the following questions:

Q1. Can a DW in OMOP CDM format support the clinical research at InCor?

Q2. How to apply DL to learn high-level concepts used for diagnosis in clinical data?

Q3. How to apply DL to learn low-level concepts used for diagnosis in medical images?

These questions are addressed in three different but interconnected studies, each handling a particular set of materials and methods that can be combined in different steps of the general learning pipeline: data collection and preprocessing; feature extraction and engineering; classification and domain knowledge analysis. The first front **F1** refers to question **Q1** and is tackled in Chapter 2; the second front **F2** refers to question **Q2** and is covered in Chapter 3; while the last front **F3** refers to question **Q3**, whose investigation is reported in Chapter 4. In the following paragraphs we provide details for each of those fronts and their hypotheses.

F1. We propose a new data warehouse to pull data from InCor's EPR database (inside the SI3 HIS), using the OMOP CDM standard, which is an international format aimed at improving research in computer-aided medical systems. This process uses a method to extract, transform and load data between the databases such that data linkage is preserved between the anonymous output and the raw medical data. By comparing cohorts obtained with the software OHDSI Atlas and the reference Pauá study, the new data warehouse with OMOP standards can retrieve the same patient cohorts and reproduce clinical studies, and also prepare raw datasets

for DL models. We departed from the following hypothesis: *an appropriate ETL process is able to produce a new database that is both consistent with the original data and in accordance to the OMOP standard, so that the new database shall support a wide range of new research initiatives within InCor.*

F2. We propose an extension to a novel CNN architecture that uses *Aggregated Transformation* and *Squeeze-and-Excite* mechanisms combined into a residual processing block. This architecture can learn to classify images based on diagnostic labels derived from microscopy analysis, and has results comparable to state-of-the-art DL models without transferring the learning from existing general-purpose architectures. At the same time, it demonstrates an ensemble with far fewer layers and parameters than previous works. The experimental evaluation with datasets ISIC, PH2, and 7-POINT demonstrates that the use of cutting-edge CNN techniques can rival previous results using just a fraction of the number of weights. This architecture works for melanoma detection (a binary problem) and skin-lesion classification (a multi-class problem) after dataset preparation. We also demonstrate the importance of the image acquisition protocol by training and testing with unrelated datasets. We departed from the following hypothesis: *by carefully designing the CNN architecture and fine-tuning specifically for the dermatoscopic images, it is possible to learn high-level diagnostic concepts and achieve state-of-the-art classification performance with much less parameters, progressively improving it similarly to what was done for larger general-purpose object recognition models.*

F3. We propose a method to detect the region of interest (RoI) in cardiac magnetic resonance images (CMR), using the raw images. The algorithm is based on convolution operations and on the use of a radial basis function to model the heart motion in the whole exam, accumulating on the time dimension. This composes a much simpler model of the heart's energy density field and is defined in a way to allow integration into a CNN architectural block. This method is added to a pipeline with the U-Net segmentation CNN proposed by Ronneberger, Fischer and Brox (2015), and tested in three public reference datasets against the canonical U-Net. The RoI detection method preprocesses the data for CNN segmentation, and can improve the U-Net segmentation quality and processing speed. We departed from the following hypothesis: *discarding similar-looking structures that lie outside the heart region (such as blood vessels and other organs) can improve heart segmentation CNNs, and this can be realized by narrowing the CNN receptive field with a heart localization method.*

1.4   Summary of motivation, objectives

Motivation: the use of HIS for decades has consolidated significant databases with potential for clinical research and decision support in health; additionally, new data processing techniques, such as DL, demonstrated expressive results in image recognition and other complex data domains. These two recent achievements provide material together (large medical

databases) and methods (large-scale machine learning algorithms) for clinical research using retrospective data, which will allow the development of more accurate predictive models, multicentric scientific collaboration, validation of clinical trials at international scale, and digital healthcare inclusion in more locations.

Problems: the data model used by HIS was mainly focused on supporting administrative and financial activities of hospitals, and seldom contains the full history of patient's records; real medical data is complex, voluminous, and lack homogeneity. Specific programs are needed to collect and prepare the data before it can be used in DL tools.

General objective: to expand the use of medical data and the ability to automated diagnosis through the integration and homogenization of the different sources of data in a HIS, and the analysis of real medical data with state-of-the-art models of machine learning known as DL, both in clinical data and medical images.

Expected results: a new ETL project from InCor data; a database of anonymized and standardized data in an internationalized model for clinical research, using data warehousing techniques; implementation and evaluation of DL in medical imaging data domains; and DL tools to extend the use of medical data at InCor.

1.5   Summary of Contributions

As our first contribution, we migrated a large clinical database, the InCor's SI3 HIS, to the OMOP CDM standard, resulting in a data warehouse following international research standards and with ability to link raw medical data such as CMR images to the patient records. Using OHDSI Atlas software over this new data warehouse, it was possible to reproduce cohort selections from the previous Pauá database with high agreement. The new data warehouse was then used to prepare data materials for several applications and studies over the course of four years, such as in Clementino *et al.* (2020), Lima *et al.* (2021) and Linhares *et al.* (2022). Our contribution lies in demonstrating that the adoption of the CDM standard both could support previous InCor activities and nurture novel studies in different topics such as databases, information visualization and AI.

The second contribution was the extension of an advanced CNN to multi-class classification. We observed that a more sophisticated preprocessing based on colorized Contrast-Limited Adaptive Histogram Equalization (CLAHE) was able to improve the AUC and the Sensitivity by significant amounts considering the classification metrics. We also verified significant results in comparison to much larger transfer-learning-based architectures using more advanced architectural techniques. The CNN architecture is one order of magnitude smaller than previous works, indicating that it is possible to train, from scratch, a small network and still achieve a good overall performance in the multi-class scenario. For reference, the smallest state-of-the-art networks in the ImageNet dataset, including the architectures engineered by AutoML methods,

have, at least, 5 million parameters. Overall, this contribution confirmed that DL can learn high-level diagnostic concepts from raw images, and also lead to interesting discoveries.

The third contribution was Full Motion Focus (FMF), a novel approach based on convolution operations and on the use of a radial basis function that detects the RoI in CMR images. We validated FMF with a U-Net CNN comparing our results to those of the canonical U-Net and of the FMF-CNN in three public reference datasets. According to our results, FMF was able to recall 99.69% of the RoI voxels in all the datasets, being suitable to preprocess the data for CNN segmentation. FMF accelerated the training process by 150%, and also increased Sørensen-Dice coefficient in the majority of our test cases. This contribution confirms that DL can learn low-level concepts such as heart walls and motion, but also reveals that some models benefit from preprocessing techniques, specially in data with noise and artifacts.

1.6 Document Organization

In order to describe our results in detail, this thesis is organized as a Collection of Articles divided into five chapters, in which, besides the Introduction and Conclusion chapters, the intermediary ones reproduce selected articles resulting from the thesis project. The document organization and the articles contained in each chapter is as follows:

- Chapter 1 introduces the problem, context, and motivation underneath the thesis;

- Chapter 2 presents the classical database analysis framework, collecting structured and semi-structured electronic patient records for transformation to a data warehouse;

  It covers the following article:

  Lima *et al.* (2019). *Transforming two decades of EPR data to OMOP CDM for clinical research.* Studies in Health Technology and Informatics 264, IOS Press.

- Chapter 3 presents a deep learning application in Dermatology, which detects and classifies multiple types of cancerous skin lesions in dermatoscopic images;

  It covers the following article:

  Lima *et al.* (2021). *DermaDL: Advanced convolutional neural networks for computer-aided skin-lesion classification.* SN Computer Science 2(253), Springer Nature.

- Chapter 4 presents a deep learning module in Cardiology, which helps to analyze images of the heart by delineating and measuring structures in magnetic resonance imaging;

  It covers the following article:

  Lima *et al.* (2022). *Full Motion Focus: Convolutional Module for Improved Left Ventricle Segmentation over 4D MRI.* In: Proceedings of Image Analysis and Processing – ICIAP 2022.

- Chapter presents the conclusions and final remarks.

CLINICAL DATA WAREHOUSES

In this chapter, we reproduce the following article:

Lima *et al.* (2019). *Transforming Two Decades of EPR Data to OMOP CDM for Clinical Research.* Studies in Health Technology and Informatics, IOS Press. Presented in MEDINFO 2019, Lyon, France.

This article presented the extract-transform-and-load (ETL) process from the Electronic Patient Records (EPR) at the Heart Institute (InCor) to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format. We described the initial database characterization, relational source mappings, selection filters, data transformations and patient de-identification using the open-source OHDSI tools and SQL scripts. We evaluated the resulting InCor-CDM database by recreating the same patient cohort from a previous reference study (over the original data source) and comparing the cohorts' descriptive statistics and inclusion reports.

The results exhibited that up to 91% of the reference patients were retrieved by our method from the EPR through InCor-CDM, with AUC=0.938. The results indicate that the method that we employed was able to produce a new database that was both consistent with the original data and in accordance to the OMOP CDM standard. The main result of this study is a curated clinical data warehouse that serves as a basis for understanding a person's health, the possible outcomes and treatments.

Generally, clinical data warehouses are curated according to statistical models of human anatomy and laboratory specimens. For example, when too much sugar is diabetes? Cobas *et al.* (2022) recommend a test for establishing a positive type-2 *diabetes mellitus* (DM2) diagnostic: i) Serum glucose on fasting $\geq$ 126 mg/dℓ, **AND** ii) Serum glucose $\geq$ 200 mg/dℓ **OR** Glycated hemoglobin HbA$_{1c}$ $\geq$ 6.5%, after 2 hours of 75 g glucose overload. In this example, the real data would be the glucose and HBA$_{1c}$ concentrations measured from each patient's blood samples. The clinical information is defined by consolidating all medical examinations and

laboratory tests in a decisive diagnostic, e.g., if the physician also observed swollen feet or other circulatory problems during the routine check-up of our hypothetical patient, she could file an ICD-10 (World Health Organization, 2016) diagnostic code E11.51–"*Type 2 diabetes mellitus with diabetic peripheral angiopathy without gangrene*" in the patient record.

The data warehouse stores both the detailed laboratory tests as raw data and the consolidated diagnostic as clinical information. In clinical studies, researchers can select patients in groups called *cohorts* with specific conditions, diseases or treatments. Then, researchers associate and correlate each combination of medical data to outcomes and diagnostics, understanding which diseases and conditions influence each other over time. This is appropriately fulfilled by an analytic data model, such as the clinical data warehouse described in the article.

Clinical data warehouses can also prepare datasets for other tasks (such as deep learning) by linking raw data to many sources of information in the data warehouse, then formatting in a computable format such as dense arrays or knowledge graph trees. Clementino *et al.* (2020) demonstrate a derived study that takes advantage of the OMOP CDM format in its cohort selector, and then transforms the data to a vector space model for clustering.

In the following pages, we reproduce the article from Lima *et al.* (2019). *This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*

Transforming Two Decades of ePR Data to OMOP CDM for Clinical Research

Daniel M. Lima[a], Jose F. Rodrigues-Jr[a], Agma J. M. Traina[a], Fabio A. Pires[a], Marco A. Gutierrez[b]

[a] *Institute of Mathematical and Computer Sciences (ICMC), University of São Paulo, São Carlos, São Paulo, Brazil,*
[b] *Heart Institute (InCor), Clinics Hospital, Faculty of Medicine, University of São Paulo, São Paulo, São Paulo, Brazil*

Abstract

This paper presents the extract-transform-and-load (ETL) process from the Electronic Patient Records (ePR) at the Heart Institute (InCor) to the OMOP Common Data Model (CDM) format. We describe the initial database characterization, relational source mappings, selection filters, data transformations and patient de-identification using the open-source OHDSI tools and SQL scripts. We evaluate the resulting InCor-CDM database by recreating the same patient cohort from a previous reference study (over the original data source) and comparing the cohorts' descriptive statistics and inclusion reports. The results exhibit that up to 91% of the reference patients were retrieved by our method from the ePR through InCor-CDM, with AUC=0.938. The results indicate that the method that we employed was able to produce a new database that was both consistent with the original data and in accordance to the OMOP CDM standard.

Keywords:

Medical Informatics; Health Information Exchange; Data Curation

Introduction

In the last decades, the field of Informatics has unveiled the, so-called, Big Data phenomenon; an intense increase of data gathering, exchanging and storage in several human activities. This phenomenon is outlined by the so called five Vs: velocity, volume, variety, veracity and value of data; posed as the big challenges for data analysis and processing [1]. Such characteristics are also common in Medical and Health Information Systems, whose databases have grown into huge amounts of patient information and health-related activities, in diverse formats, always online, and easily accessible from a mobile screen. If properly interconnected and treated, these systems offer interesting data sources for evidence-based research, such as Precision Healthcare, Population Health, Clinical Research, and more. [11]

However, this data exchange is often a significant challenge. Most of the Electronic Patient Record (ePR) systems were not explicitly designed for research; rather, they are organized by standards and structures which are local to the institutions they primarily serve, e.g. hospital facilities, clinics, pharmacies, health insurance companies, etc. Thus, interchange methods, protocols and architectures were designed to cope with this challenge, such as the HL7 standards [12] and the OpenEHR platform [13]. These approaches mediate the communication of near-line and online transaction processing (OLTP) systems, specifying standard data elements and transformations from their internal data to a common messaging format. Furthermore, data analysis follows an approach akin to online analytical processing (OLAP), using denormalized, coalesced and preprocessed data in a standard common database format.

OHDSI and the Common Data Model

In this context, the Observational Health Data Sciences and Informatics (OHDSI – www.ohdsi.org) initiative grew out of the Observational Medical Outcomes Partnership (OMOP) developing a mature data standardization model, the OMOP Common Data Model (CDM) [6]. Having a ready-to-use database in a standard common model such as the OMOP CDM simplifies the exchange and integration of standardized methods, applications, information and tools between clinical researchers; a critical feature for distributed research networks using patient-centric clinical databases. [7]

The CDM is a strong information model; its conceptual elements and their relationships are explicitly specified in a formal language, and every piece of information is connected to a standard term from SNOMED-CT. The CDM's Standard Clinical Tables include Person, Visits, Observations, Conditions, Death, Procedure occurrences, Drug exposures, Measurements and more detailed information such as Drug ingredients, and Condition modifiers. OHDSI also provides open-source CDM applications for visualization and statistical analysis of patient-exposure-outcome cohorts. [10]

The InCor data integration challenges

The Heart Institute (InCor) of São Paulo, Brazil, is one of the six institutes of the Clinics Hospital complex, University of São Paulo Medical School. In the last two decades, InCor has increased its commitment in integrating all the relevant information of its patients, successfully developing an ePR named SI³. The first version of SI was deployed in year 2000; currently, it stores the clinical history, examinations, procedures, surgeries, notes, laboratory tests, medication, bills, and more for 1.3 million patients. Since then, the system has continuously evolved, overcoming several challenges related to the exchange of information among different healthcare institutions and remote installations. Furuie et al., describe details of the system architecture [8], while a number of studies involved cohort selection based on information collected by the SI³ system [3-5].

However, the workload involved in extracting the relevant patient information from SI³ has motivated the adoption of new strategies. Recently, we started the mapping from the SI³ data model to a standard data model that can simplify the observational retrospective studies related to Clinical Research. In this paper, we present the steps related to the mapping between SI³ and the CDM data models to prepare a new standardized database, named InCor-CDM, that can be used with

the OHDSI toolset and a number of visual analytics tools. We measure the quality of the resulting InCor-CDM database using precision and recall statistics, when compared to the cohort generated by a previous study (gold standard).

Methods

Environment preparation

We prepared the InCor-CDM database environment by installing a PostgreSQL 10 DBMS, Java 10 JDK and Docker-compose in a Linux workstation. The database setup includes: (a) creating the required database users; (b) creating the OMOP CDM tables with the CommonDataModel/PostgreSQL scripts; and (c) importing the standard OMOP vocabularies from athena.ohdsi.org. All the OHDSI sources are available at github.org/OHDSI. Next, we installed the Achilles and Broadsea repositories required for the OHDSI web applications, configured the addresses and JDBC URLs and started their respective docker containers.

Database characterization

The InCor SI³ database is stored in an Oracle 12c instance, accessed with the Oracle JDBC connector and DBeaver SQL client. The first step to manage the database was sampling its tables and columns with the OHDSI's WhiteRabbit software. This application generates a spreadsheet with the most frequent values of each selected column, so we can inspect them and make decisions about which columns to ignore, especially those with irrelevant or missing values. Then, we use RabbitInAHat to parse the output of WhiteRabbit and draw relational data-flow diagrams for documentation.

Some namespaces in SI³ were ported from non-relational or older systems; such namespaces had no associated documentation, comments, constraints or foreign keys. For this reason, the database metadata had to be converted to a searchable JSON format with table and column names, types, comments, constraints and foreign keys. The column names were tokenized according to the naming scheme of the institution (e.g. abbreviations separated by underscore) and matched to similar columns in other tables; the goal was to find implicit relationships where the foreign keys were missing. Some attribute domains, such as internal record status codes, event sequence and timing diagrams, were documented from interviews with the support staff of InCor.

Patient de-identification

For de-identification purposes, personal information mapped to the CDM was limited to a minimum. Any key with a path to a patient primary key (and the PK itself) was **pseudonymized** [14], i.e., direct identification information such as citizenship document, phones numbers, addresses and names are not ported to the CDM, and the record primary key is exchanged to a *new id* (pseudonym), which is a random number drawn from an uniform distribution in the range 1×10^{10} and 9×10^{15} using Oracle's DBMS_RANDOM functions, addressing collisions with repeated sampling. This range was selected not to conflict with OMOP's standard concept ids (0 up to 2×10^9 are reserved) and to be within the limits of JSON numbers (53-bit precision). InCor holds the mapping from the new ids to the original keys in a private table; the mapping is to be used for notifying the patient, or her/his physician, in case the result of a study can improve a patient's condition.

Numerical variables were truncated in order to satisfy a baseline level of **k-anonymity** [15], i.e., guaranteeing that any patient variable value have at least *k* patients with the same

information, so no patient is uniquely identifiable. For example, event dates were truncated to yearly, monthly, daily or hourly precision where original precision were not needed. Records with spurious attributes (e.g., dates in the future, outside any visit, invalid range, null required field, missing keys) were discarded. Also, we only loaded data from patients born before 2010 (aged 18+), with at least one valid visit.

si3.pac_paciente

paci_id	tp_sexo	...
01721	M	...
01722	M	...
01723	F	...

keys

table	src_id	new_id
pac_paciente	01721	7369111123
pac_paciente	01722	1257321234
pac_paciente	01723	3038618654

gender_map

src	id
F	8532
M	8507

omop.person

person_id	gender_concept_id	...
7369111123	8507	...
1257321234	8507	...
3038618654	8532	...

Figure 1– Example: gender mapping from SI³ to the CDM.

Figure 1 shows sample data to illustrate this process. Each column from CDM person table (e.g., gender_concept_id) is extracted from the *source* table and column (e.g., pac_paciente. tp_sexo), and transformed with the appropriate domain map (e.g., value 'F' used in SI³ for female gender is mapped to CDM concept 8532). The *keys* and *gender_map* tables are populated beforehand. The standard CDM concept ids were searched in OHDSI's Athena – a web-based CDM vocabulary explorer. Then, observe the following query, which loads data from the SI³ PAC_PACIENTE table into the InCor-CDM omop.person table with remapped keys and concepts:

```
INSERT INTO omop.person
SELECT K.new_id          AS person_id,
       EXTRACT(YEAR FROM P.dt_nasc)
                         AS year_of_birth,
       COALESCE(G.id, 0) AS gender_concept_id
FROM si3.pac_paciente P
JOIN keys K ON  K.table='pac_paciente'
           AND K.src_id=P.paci_id
LEFT JOIN gender_map G ON P.tp_sexo=G.src;
```

This query operates on the table samples in Figure 1, where it is assumed that the *keys* table holds the patient's random new_id, and that the *gender_map* table is a domain map table defined as (src char(1); id integer) corresponding to values of (tp_sexo, gender_concept_id). Unmapped values receive code 0, meaning "unknown concept" in the standard CDM vocabulary. Related tables were joined to the patient PK as usual, with their PK also remapped by keys.new_id.

Coding translation

InCor SI³ uses the ICD 10 for diagnosis, a set of Brazilian vocabularies for coding clinical conditions, drugs, and procedures (TUSS is used by the Brazilian Health Care System for general terms and Brasindice for drugs – datasus.gov.br) and an internal coding system for generic billable items. Internal codes in use at InCor were inserted as new Concepts in the CDM (with ids mapped between 3×10^9 and 9×10^9) under the "InCor" vocabulary, with Concept Relationships to standard concepts whenever this information was available in SI³. Initially, the records are inserted in the InCor-CDM with the original source codes, then the Concept Relationships from local to international codes are used to update the local InCor-CDM references to OMOP standardized terminologies, such as the SNOMED-CT,

RxNorm, and LOINC [6], while the preserving the original source code stored in the patient record.

Quality assessment

After loading all the CDM tables, we execute the Achilles analysis, which will report data quality issues as errors, warnings or notifications. It will also preprocess the demographic characterization of the database for the visualizations and reports; after that, we start the Atlas WebAPI and front-end servers. Atlas offers both a RESTful API and a graphical web interface to schedule the execution of OHDSI methods, automatically generating the queries on a properly built and configured CDM database.

The most relevant errors initially found were related to: (a) the referential integrity, e.g., events without associated visit or events without a valid person_id, which were discarded; (b) lack of condition_eras, drug_eras and observation_periods, which were imputed from the ePR records by collapsing all the events of a patient not apart for more than a year into a single era – in the CDM standard, an "event" era refers to a time period of interest where "events" are recorded in the ePR; (c) events with invalid date, e.g. a condition_occurrence with start date in a future date; (d) too many patients without a diagnostic or prescription; InCor-CDM keeps such records because they can be used in the condition or drug_exposure dashboards, regardless of previous diagnosis.

Statistical evaluation

After initial corrections, we evaluate the quality of InCor-CDM by using software OHDSI's Atlas to recreate a previous CVD patient cohort observed in a reference study that was executed over the InCor SI³ database by Abrahao et al. [9]. That reference study prepared a clean de-identified database named Pauá, based on a 2016 snapshot of SI³ records of patients, admissions, discharges, diagnoses, surgeries, PCI, medications, and laboratory tests. Within the Pauá database, Abrahao et al., verified the effect of statins on the survival rate of patients diagnosed with cardiovascular diseases. Our study will evaluate the InCor-CDM quality by defining a cohort with the same criteria and by verifying how many patients (with the same private InCor identifiers) were retrieved by each criteria given that reference study.

Results

In this section we evaluate InCor-CDM quality by selecting the same DCV cohort of a previous study in the Pauá database [9], by using Atlas over the InCor-CDM database. Our evaluation computes the Area Under the ROC Curve (AUC) based on the results of 12 executions of the DCV cohort at various settings. We compute predictive statistics for each cohort execution, using Pauá as a gold standard. The resulting AUC ranges from 0.5 (no different than random sampling) to 1.0 (reproduces exactly the same result as the gold standard).

Table 1– Databases cardinality (thousands of records).

Domain	SI³2016	Pauá	SI³2018	CDM
Person	1,116	323	1,346	946
Visit Occurrence	6,427	5,686	7,499	7,305
Condition Occur.	1,205	1,007	1,361	1,324
Procedure Occur.	45,024	144	53,945	51,479
Drug Exposure	83,283	2,775	100,052	38,962
Measurement	22,025	20,528	31,095	30,177
Death	17	21	18	18

In Table 1, we present the cardinality of the InCor-CDM in comparison to the SI³ and Pauá database. Note that the Procedure and Drug domains in the Pauá database have substantially less records; this is because they are restricted to surgeries and particular classes of drugs, and because Pauá only uses patients with at least one admission and diagnosis. Additionally, InCor-CDM is based on a more recent snapshot, named SI³2018 in the aforementioned table.

Cohort definition

To replicate the Pauá study, we created a cohort over the InCor-CDM database with the criteria below. Each list item directly corresponds to a HTML input field in the Atlas **cohort definition form**. These criteria are translated to SQL queries over our database in the CDM format (InCor-CDM).

- *Initial Event Cohort*: People having any of a **visit** occurrence of **Outpatient concept set** (with concept_id 9202, Outpatient Visit);

- *Additional Qualifying Inclusion Criteria*:
  - *Condition occurrence criteria*: with **at least 1** of **any condition**;
  - *Demographic criteria*: **age** greater than or equal **23** (in the censor window), matching the patients over 18 years old at the Pauá study start (1999);
  - *Demographic criteria*: with a **gender** of **MALE** (8507) or **FEMALE** (8532);
  - *Condition occurrence criteria*: of **CardioVascular Disease** (a concept set of concept ids from ICD-10 categories I20 to I25, I64 to I70 and G45, including descendants and mapped), with occurrence start between **2003-01-01** and **2013-12-31**, where event starts between **All** days before and **30** days after index end date (meaning the diagnosis was recorded around the time of the initial outpatient visit);
  - *Visit occurrence criteria*: with **at least 1** of **Outpatient concept set**, where event starts between **30** after and **All** days after index end date (a subsequent visit recording the outcome);
  - *Limit qualifying cohort to the* **earliest** *event*;

- *Era collapse gap size*: **1 day**;

- *Cohort censor window*: starting **2003-01-01** and ending **2013-12-31**.

After verifying the Concept Sets to have the correct concept ids, we generated the cohort on the InCor-CDM database and verified the cohort attrition report (the number of patients remaining in the selection after each filter) in Table 2:

Table 2– InCor-CDM CVD cohort attrition report in Atlas.

Criteria	n	%	Visualization
Initial	778,015	100.00	
a) Dx	351,205	45.14	
b) 18+	321,827	41.37	
c) M/F	303,847	39.05	
d) CVD	45,710	5.88	
e) 2nd-V	39,910	5.13	
f) People	39,498		

Table 2 summarizes the cohort definition criteria, where each abbreviation mean: (a) **Dx**: has a valid condition occurrence (diagnosis); (b) **18+**: over 18 years old; (c) **M/F**: Male or Female gender; (d) **CVD**: has any occurrence of ICD-10 I20 to I25, I64 to I70 or G45; (e) **2nd-V**: has a second outpatient visit occurrence more than 30 days after the index event (the initial visit); (f) **People**: lists how many people actually matched the events, because some patients have more than one episode. The baseline results achieved by the InCor-CDM CVD cohort indicate that 39,498 patients satisfy the criteria.

Table 3– Pauá reference CVD cohort attrition report.

Criteria	n	%	Visualization
Initial	313,894	100.00	
Dx	313,894	100.00	
18+	282,677	90.00	
M/F	263,339	83.87	
CVD	56,799	18.06	
2nd-V	27,698	8.80	
People	27,698		

In comparison, Table 3 displays an attrition report for the reference study (Pauá). The visualization was drawn with Python and GNU Gimp because Pauá's schema is not based on the CDM standard, and thus could not be used in Atlas. The reference cohort selected with the same criteria in the Pauá database has 27,698 patients (30% less than our first result), indicating the need of further refinement.

Initial Evaluation

All subject_ids in the cohort defined in Atlas were compared to the set of patient ids in the Pauá cohort. This comparison was executed by computing confusion matrices against the Pauá cohort result as a gold standard, then the derived scores: true and false positive ratios (TPR/FPR), positive and negative predictive power (PPV/NPV), accuracy (ACC) and F1-score. The confusion matrix the cohort #1 is given in Table 7, and derived scores make the first row of Table 8.

Table 7– Confusion matrix for cohort #1 (in Table 2).

Pauá InCor-CDM	P	N	Total
P	25,423	12,290	37,713
N	2,651	282,887	285,538
Total	28,074	295,177	323,251

Parameter refinement

Therefore, we noted cohort accuracy variations by creating additional cohorts with slightly adjusted parameters for the CVD criteria. Table 4 displays the attrition reports with increased periods for the qualifying criteria of CVD occurrence start date after index end date. Table 5 displays the attrition reports with increased periods for the qualifying criteria of Cardiopathy occurrence start date. We observe that increasing the collapse gap size to 7 and 14 days had no substantial effect on the results (Table 4).

Table 4– Varying condition start after index event (days).

Criteria \ days	7	14	21	30
Initial		778,015		
Dx, 18+, M/F		303,847		
CVD	44,967	45,255	45,484	45,710
People	39,055	39,203	39,342	39,498

Table 5– Varying condition start periods (years).

Criteria \ days	2003-2013	2000-2013	2000-2016
Initial		778,015	
Dx, 18+, M/F		303,847	
CVD	45,710	49,942	63,656
People	39,498	43,293	54,126

Table 6– Varying 2nd visit event start after index (days).

Criteria \ days	All	365	180	90
Initial		778,015		
Dx, 18+, M/F		303,847		
CVD	45,710	44,228	43,950	43,667
People	39,498	35,457	32,767	29,414

Then we evaluated all the patients selected in each cohort (from Tables 4, 5 and 6) by comparing them to the reference Pauá study, using the private Keys table to map the CDM person_ids to the SI³ patient ids (confidential to the institution), whose results are presented in Table 8. It should be noted that Pauá had an update in October 2016, after the reference study was published, and so we re-executed the query for its patient cohort, resulting in P=28,074 patients selected in the cohort (1.4% increase) and overall total P+N=323,251 patients included (3.0% increase).

Table 8– Predictive scores for each cohort.

#	TPR	FPR	PPV	NPV	ACC	F1
1	.905	.041	.674	**.990**	.953	.772
2	.901	.040	.678	**.990**	.954	.774
3	.903	.040	.677	**.990**	.954	.774
4	.904	.041	.676	**.990**	.954	.773
5	.905	.041	.674	**.990**	.953	.772
6	**.907**	.052	.623	**.990**	.944	.738
7	**.907**	.052	.622	**.990**	.944	.738
8	**.907**	.052	.620	**.990**	.943	.736
9	.889	.040	.680	**.990**	.954	.775
10	.877	.031	.727	.988	**.960**	**.795**
11	.829	.027	.743	.983	**.960**	.784
12	.754	**.023**	**.752**	.976	.957	.753

Evaluation of the ROC curve

We complete the evaluation by plotting the ROC curve from the predictive scores in Table 8. Only the patients existing in the Pauá database were used, i.e., only 37,713 patients of those retrieved in the cohort #1 also existed in the Pauá database. The highest scores of each column are highlighted in bold, e.g., cohort #10 exhibited the highest F1-score.
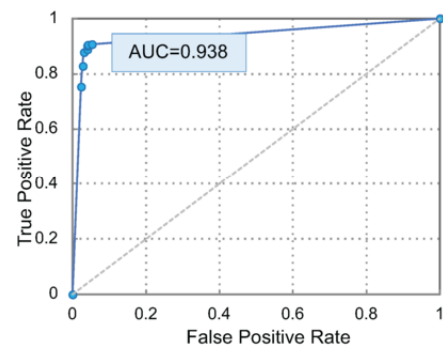

Figure 2– Empirical ROC curve for Table 8.

In this section we presented the results of the cohorts selected with Atlas from the InCor-CDM database. We also presented the effect of tweaking cohort parameters in the quality of the results (Table 8), which exhibited up to 80% F1-score, 75% Precision (PPV) and 91% Recall (TPR) at different settings. Figure 2 summarizes our results in an empirical ROC curve, exhibiting 0.938 of area under curve (AUC).

Discussion

The resulting InCor-CDM exhibits high agreement with the previous gold standard study [9]. This means that it is possible to estimate the same population-level effects (e.g. different medications) in both databases. InCor-CDM additionally benefits from the quality analyses implemented in the OHDSI Achilles tool, which warns about inconsistencies and errors found in the transformed data, and can be used for more advanced analysis in comparison with external CDM-based databases. For future work, we envision the comparison of patient cohorts between InCor-CDM and external CDM-based databases from other OHDSI work groups, further studying data quality, subpopulation characteristics between different institutes and evaluating risk scores for InCor patients.

Conclusions

We presented details of the migration process of a huge clinical database, the InCor's SI^3, to the OMOP CDM standard, an international format aimed at improving research in computer-aided medical systems. We presented the method used to extract, transform and load data between the databases commenting on the challenges regarding models, formats, terminology, and tools. We evaluated the quality of the resulting database, named InCor-CDM, by comparing cohorts obtained with the software OHDSI Atlas. We considered a previous cohort selection study used as ground truth; for a systematic comparison, we computed several information retrieval statistics and a ROC curve. The cohorts defined in Atlas exhibited from 62% to 75% precision, 75% to 91% of recall, 74% to 80% F1-score, and 0.938 of area under the ROC curve (AUC). The results indicate that the method that we employed was able to produce a new database that was both consistent with the original data and in accordance to the OMOP standard. The new database shall support a wide range of new research initiatives within the Heart Institute.

Acknowledgements

References

[1] V.N. Gudivada, R. Baeza-Yates, and V.J. Raghavan, Big Data: Promises and Problems, *IEEE Computer* **48, 3** (2015), 20-23.

[2] A.K. Jha, D. Doolan, D. Grandt, T. Scott, and D.W. Bates, The use of health information technology in seven nations, *Int J Med Inform* **77, 12** (2008), 848-854.

[3] M.T. Abrahão, M.R.C. Nobre, and M.A. Gutierrez, Descriptive Statistics of 65000 Patients Treated for Myocar-

dial Ischemia: data from routine electronic health records, *Evidence Live* **13** (2013).

[4] M.T. Abrahão, M.R.C. Nobre, and M.A. Gutierrez, The effectiveness of statins in the treatment of cardiovascular disease: Cross-sectional study with paired groups from electronic patient records, *Value Health* **16, 7** (2013), A518.

[5] M.T. Abrahão, T.J. Soares, F.A. Pires, M.A. Gutierrez, and M.R.C. Nobre, Data Warehouse com dados da Saúde Pública: Estudo de Caso sobre o Tratamento de Doença Cardiovascular Aterosclerótica no Estado de São Paulo, *CBIS* **12** (2010).

[6] G. Hripcsak, J.D. Duke N.H. Shah, C.G. Reich, V. Huser, M.J. Schuemie, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *MedInfo* **2015** (2015), 574-578.

[7] D. Madigan, P.B. Ryan, M. Schuemie, P.E. Stang, J.M. Overhage, A.G. Hartzema, et al. Evaluating the Impact of Database Heterogeneity on Observational Study Results. *Am J Epidemiol* **178, 4** (2013), 645-651.

[8] S.S. Furuie, M.F.S. Rebelo, R.A. Moreno, M. Santos, N. Bertozzo, G.H.M.B. Motta, et al. Managing Medical Images and Clinical Information: InCor's Experience, *IEEE Trans Inf Technol Biomed* **11, 1** (2007), 17-24.

[9] M.T. Abrahão, M.R.C. Nobre, and M.A. Gutierrez, A method for cohort selection of cardiovascular disease records from an electronic health record system, *Int J Med Inform* **102** (2017), 138-149.

[10] D. Yoon, E.K. Ahn, M.Y. Park, S.Y. Cho, P. Ryan, M.J. Schuemie, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research, *Healthc Inform Res* **22, 1** (2016), 54-58.

[11] E.A. Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M.J. Schuemie, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc* **22, 3** (2015), 553-564.

[12] R.H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron, and A. Shabo, HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* **13, 1** (2006), 30-39.

[13] D. Kalra, T. Beale, and S. Heard, The openEHR foundation, *Studies in Health Technol & Inform* **115** (2005), 153-173.

[14] K. Pommerening, and M. Reng, Secondary use of the EHR via pseudonymisation. *Studies in Health Technol & Inform* **103** (2004), 441-446.

[15] R.J. Bayardo, and R. Agrawal, Data privacy through optimal k-anonymization, *ICDE* **21** (2005), 217-228.

Address for correspondence

Daniel Mário de Lima <danielm@usp.br> (+55 11 2661 5545)

3

SKIN LESION IDENTIFICATION

In this chapter, we reproduce the following article:

Lima *et al.* (2021). *DermaDL: Advanced Convolutional Neural Networks for Computer-Aided Skin-Lesion Classification.* SN Computer Science, Springer Nature.

The article contemplates an application for detection and classification of skin lesions using dermatoscopic images. It extended the article of Rodrigues, Brandoli and Amer-Yahia (2020) to multi-class classification, including a colorized adaptive histogram equalization, and using `tf.keras` API. The neural network could then detect several types of lesions at once.

Early identification of the type of skin lesion is of paramount importance, because some types are carcinogenic and may become difficult to treat after a few months. Skin cancer is the most common of all cancers, and some types can spread to other organs. Instituto Nacional de Câncer do Brasil (2020) expected around 185 thousand new skin cancer cases per year during 2020-2022 in Brazil (177 thousand non-melanoma cases and 8,450 melanoma cases), while American Cancer Society (2022) estimated around 109 thousand new melanoma cases by the end of 2022 in the United States (non-melanoma cases are not required to be reported there).

The main recommendations are UV protection and routine examinations with the dermatologist, who decides to proceed with a biopsy when there is any signal of tumor development. The routine examination involves applying a liquid or gel in each skin abnormality (moles, spots, scars and others) and taking pictures with a imaging device called dermatoscope, which embeds special software for automated analysis.

Currently, the use of Convolutional Neural Networks (CNNs) is the mainline of investigation for the automated analysis of such lesions. Most of the existing works, however, were designed by transfer learning general-purpose CNN architectures, adapting existing methods trained for ImageNet and other generic photo archives to the domain of dermatology. Despite its effectiveness, this approach poses inflexibility and high processing costs.

In the article, we employed a novel architecture that benefits from cutting-edge CNN techniques Aggregated Transformations combined to the mechanism of Squeeze-and-Excite organized in a residual block. Those techniques are general formulations of previous architectures with skip connections and basic attention mechanisms. The architecture was adapted and trained from scratch to solve both the binary melanoma detection problem, as well as the multi-class skin lesion classification problem. We also used cross-dataset training/validation due to concerns about data leakage after the augmentation step.

Our results demonstrated that such an architecture is competitive to major state-of-the-art architectures adapted to the domain of skin-lesion diagnosis. The architecture is prone to evolve and to provide low processing cost for real-world in situ applications using a much smaller number of weights if compared to previous works. The processing cost specially benefits applications that run neural network inference in energy-efficient portable devices.

ORIGINAL RESEARCH

DermaDL: Advanced Convolutional Neural Networks for Computer-Aided Skin-Lesion Classification

Daniel M. Lima¹ · Jose F. Rodrigues-Jr¹ · Bruno Brandoli² · Lorraine Goeuriot³ · Sihem Amer-Yahia³

Abstract

Early identification of the type of skin lesion, some of them carcinogenic, is of paramount importance. Currently, the use of Convolutional Neural Networks (CNNs) is the mainline of investigation for the automated analysis of such lesions. Most of the existing works, however, were designed by transfer learning general-purpose CNN architectures, adapting existing methods to the domain of dermatology. Despite effective, this approach poses inflexibility and high processing costs. In this work, we introduce a novel architecture that benefits from cutting-edge CNN techniques Aggregated Transformations combined to the mechanism of Squeeze-and-Excite organized in a residual block; our architecture is designed and trained from scratch to solve both the binary melanoma detection problem, as well as the multi-class skin-lesion classification problem. Our results demonstrate that such an architecture is competitive to major state-of-the-art architectures adapted to the domain of skin-lesion diagnosis. Our architecture is prone to evolve and to provide low processing cost for real-world in situ applications using a much smaller number of weights if compared to previous works.

Keywords Deep learning · Dermatology · Skin lesion classification · Melanoma detection · Skin cancer

Introduction

During the early stages, it is difficult to tell apart malignant skin lesion from certain benign forms of skin cell differentiation; both kinds of lesion look similar and pose challenges to visual inspection [1]. According to a study carried out

This article is part of the topical collection "AI and Deep Learning Trends in Healthcare" guest edited by KC Santosh, Paolo Soda and Zalelam Temesgen.

✉ Jose F. Rodrigues-Jr
    junio@usp.br

    Daniel M. Lima
    danielm@usp.br

    Bruno Brandoli
    brunobrandoli@dal.ca

    Lorraine Goeuriot
    lorraine.goeuriot@imag.fr

    Sihem Amer-Yahia
    sihem.amer-yahia@imag.fr

¹ University of Sao Paulo, Sao Carlos, SP, Brazil

² Dalhousie University, Halifax, Nova Scotia, Canada

³ CNRS, Universite Grenoble-Alpes, Grenoble, France

by Menzies et al. [2], general practitioners had a sensitivity and specificity for detection of melanoma of 62% and 63%; dermatologists had a sensitivity and specificity of 80% and 60%. These facts indicate that, clinically, it is difficult to differentiate the early stage of malignant melanoma from certain benign skin lesions due to their similarity in appearance. Therefore, it is desirable to have computer-aided support in analyzing skin lesions; as we will demonstrate, this support has the potential to overcome reported human-based sensitivity and specificity rates making it a trustworthy second opinion.

The most severe type of skin lesion is melanoma, it has the potential to evolve and disseminate malign tumors over the body, particularly in the lungs and brain. In the USA alone, according to the statistics of the American Cancer Society [3], in 2019, 96,480 new melanomas are to be diagnosed with roughly 7,230 casualties. In the world, nearly 50,000 deaths occur annually [4]; it is the fifth most common kind of cancer. As for any kind of health problem, early detection is key for a high rate of survival. Similarly, carcinoma, a more broad type of cancer, can originate from differentiated skin cells, posing severe health threats [5]. Meanwhile, in the last decade, deep CNNs (Convolutional Neural Networks) have achieved unprecedented

levels of performance in computer vision tasks; this is due to four factors: the explosion in data production; the popularization of high-density parallel hardware in the form of general-purpose computing on graphic processor units (GPGPU); algorithmic improvements on artificial neural networks; and the advent of large-scale linear algebra frameworks.

Many works have addressed the task of automatic screening of skin lesions via dermatoscopy images, with deep convolutional networks being the leading technology. Digital images of skin lesions have the potential for educating the general public and for supporting health providers in early detection, as well as directly supporting remote diagnosis through teledermatology, clinical decision support, and automated risk assessment. However, the main works on automatic screening of skin lesions [6–14] have appealed to transfer the neural learning from architectures trained over image recognition tasks that, originally, did not include skin lesions—this process is known as transfer learning [15]. Although the results of this approach have demonstrated effective, it has a few drawbacks: (i) the layers of the CNN architecture responsible for features extraction cannot be altered, reducing the flexibility of the designer in improving the performance or computational cost of the network; and (ii) training the last layer on the new skin image data carries restrictions derived from the original data training, especially with respect to the data distribution, which must be known. Furthermore, learning parameters from real skin imaging datasets enables tailored neural architectures to potentially achieve higher accuracy. Differently, having the entire process of neural network engineered for the specific goal of skin-lesion classification shall provide more freedom, design opportunities, and control over the underlying mechanisms. This work proposes one such process.

As mentioned, the state-of-the-art relies on transfer learning from general-purpose image recognition architectures. Differently, we train on a dataset exclusively made of skin lesions, which grants a specialized training cycle; and also, an architecture that permits skin-lesion classification to advance on its own, instead of relying on general-purpose ones pre-trained for object recognition. Our contribution demonstrated that it is possible to engineer a specialized network prone to evolve more specifically, possibly reproducing the advances observed in the ImageNet Large Scale Visual Recognition Challenge [16]. In fact, there is room for a less computationally costly architecture capable of leveraging functionalities specific to skin lesions, like tracking asymmetry and border irregularity. In this line of work, we reached an architecture with fewer layers and relying on the architectural mechanism of *Aggregated Transformations* [17] combined to the mechanism of *Squeeze-and-Excite* [18], which resulted in an architecture with fewer weights, but similar performance. Furthermore, we conducted experiments relying on datasets from different sources, which

permitted us to demonstrate the importance of image acquisition protocols when performing automated skin-lesion classification. As we show, our results were encouraging in terms of classification performance and architecture complexity.

We contribute by:

- providing a new architecture for the classification of skin lesions;
- putting together the concepts of Squeeze-and-Excite and Aggregated Transformations;
- comparing popular architectures in a unified setting;
- introducing an inexpensive solution that fits low-performance devices.

In the next section, we provide Background to this work. Following, in the section of Related Works, we review the state of the art, basically based on transfer-learning designs. In a section dedicated to Convolutional Neural Networks, we review concepts necessary to understand our neural network architecture. After that, in section DermaDL Methodology, we introduce our architectural design, which is experimented in the binary task of melanoma detection, and in the multi-class task of skin-lesion classification. Our concluding remarks are in the last section.

Background

Histogram Equalization

The goal of technique histogram-based equalization is to enhance the contrast of images. As a result, further image-processing techniques can, potentially, produce sharper results, depending on the method.

Let I be a given image, or m_r by m_c matrix of pixel values in the range $[0, L-1]$, where L is the number of possible intensities. Let h correspond to the normalized histogram of I with one bin for each possible value as follows:

$$h_i = \frac{\#pixels\_with\_intensity\_i}{m_r * m_c}. \tag{1}$$

Then, the histogram-based equalized image E is given by

$$E_{i,j} = \lfloor (L-1) \sum_{n=0}^{I_{i,j}} h_i) \rfloor \tag{2}$$

where $0 \le i < m_r$ and $0 \le j < m_c$.

For improved results, we apply an advanced variation of histogram equalization named CLAHE, as detailed in Sect. 5.1.

Transfer Learning

Transfer learning is a strategy in which a model developed for a task is reused as the initial stage (features extraction) to define a new model tailored to a different, but similar, task. Common in computer vision and natural language processing, its advantages come from the reuse of successful CNN architectures at the cost of higher computational loads when used in production.

Dermatological Diagnosis

Dermatoscopy is a procedure carried out by dermatologists, who visually examine skin lesions to detect anomalous features that might characterize a specific disease. This procedure uses an optical apparatus (a dermatoscope with lenses and camera) to obtain high-quality images of the skin, which are evaluated according to a set of criteria, such as the "ABCDE" [19]. This acronym summarizes the most distinguishing features of the group of diseases called *cutaneous melanoma*, i.e., a cancerous growth of pigmented cells (melanocytes) in the skin. Using convolutional neural networks, the corresponding feature extraction is able to detect asymmetry and border irregularity, traits indicated by the ABCDE criteria.

The ImageNet Challenge

The ImageNet Large Scale Visual Recognition Challenge [16] is a competition for image recognition algorithms. The ImageNet dataset contains more than a million images classified into one thousand classes; 50 k pictures are randomly sampled and released before the competition for validation, and another sample with 150 k images is used for the final scoring of the submissions. One of the challenges is to assign the top five classes most similar to each object portrayed in a given image. In 2012, the AlexNet CNN [20] achieved an error rate of 15.3% in the top-5 challenge, granting the 2012's best result and also a breakthrough for image recognition with neural networks. Afterward, at least 95% of the ImageNet submissions and 100% of the first places from 2013 to 2017 derived their algorithms from research on neural networks (http://image-net.org/). By means of transfer learning, the networks that won the ImageNet competition are massively used in the task of melanoma detection.

Related Works

Since we deal with the problem of computer-aided skin-lesion screening as a binary classification (melanoma/non-melanoma), or as a multi-class classification (melanoma, carcinoma, keratosis, nevus, and other), we separated the related works into two groups. Accordingly, we refer to these two groups when analyzing our results, either the binary problem in Sect. 6, or the multi-class problem in Sect. 7.

Melanoma Detection

The use of Deep Learning (DL) to aid the diagnosis of skin cancer and non-cancerous lesions has been pushed forward in the work of Esteva et al. [6], who not only were able to diagnose skin cancer but also to create a taxonomy of skin lesions that extends the knowledge in the field. Their approach was to transfer learn Google's Inception V3 network [21] over a very large dataset of proprietary images—to date, their work is among the most advanced results concerning computer-aided diagnosis. However, their work is not open, neither with respect to the trained network nor to their dataset, which is only partly open access. This fact makes their progress inaccessible and considering the sponsoring entities of the work, possibly, it will be of permanent restricted access. In a more recent work, Fujisawa et al. [8] demonstrates that results similar to a clinical board are achievable with datasets orders of magnitude smaller than those used in the work of Esteva et al. Similarly, Fujisawa et al. use the Inception network v3.

In 2017, during the IEEE International Symposium on Biomedical Imaging, 23 research teams applied for the challenge "Skin lesion analysis toward melanoma detection" in the category of skin cancer detection (disease classification) [22]. The winner team [9] comparatively used models VGG-16 [23], ResNet-101 [24], and Inception V3 [21], all of which originally designed for general purposes; they achieved the best result for melanoma by using the Inception network, although ResNet-101 demonstrated good results as well.

In 2018, Seog et al. [10] used the Resnet-152 network achieving an accuracy as high as 96%, however, the very authors claim that their datasets for testing and training are not versatile with respect to age. The work of Zakhem [25] draws a discussion on the use of automated skin cancer diagnosis, tracing reflections on the role of its clinical use and on the adoption of the technology by physicians. The work of Ruiz et al. [26] presents a clinical decision system that classifies images with suspicious lesions on the skin to assist as a second opinion clinical protocol. The authors combined the nearest K-neighbors method, a multi-layer Perceptron (MLP), and a parametric classifier based on Bayes' decision theory to achieve a collaborative classification system. They used a set of descriptors that comprises the variation of the sharpness, color homogeneity, mean of the R and G components, and average of the HSL color luminosity. They obtained a precision of 87.76%, inferior to the state-of-the-art works, which report results above 90%.

The work of Yu et al. [27] describes an automatic melanoma diagnosis system; the authors use a convolutional residual network to segment the images during preprocessing. For training, they use 900 pre-labeled images and 350 images for testing; this method obtained an accuracy of 85.5%. The work presented by Souza et al. [28] uses decision trees based on descriptors of asymmetry, edge, and coloring to distinguish between images that present melanoma skin cancer from others that present common spots. The highest reported accuracy was of 90%. The work presented by Aswin et al. [29] also works over the identification of melanoma using CNNs; they use a wide set of descriptors including contrast, correlation, homogeneity, and RGB color variance. Their accuracy was around 88%.

Nasr-Esfahani et al. [30] described a CNN-based melanoma diagnosis system that uses a preprocessing step to correct the illumination, to segment the images, and to enhance the images through a Gaussian filter. This method obtained an accuracy of 81%. Majtner et al. [31] proposed a system that automatically classifies melanomas; in their system, the descriptor vector of each image is obtained using RSurf, Local Binary Patterns (LBP), and Convolutional Neural Networks. Further, they use the technique Support Vector Machines (SVM) for classification. During training, they used 900 images; for testing, they used 379 images. The highest accuracy reported by the authors was of 82.6%.

Multi-class Skin-Lesion Classification

Hameed et al. [11] analyzed a total of 15,546 images from 6 public datasets (ISIC2016, DermIS, DermNZ, DermQuest, PH^2, and 11K), literature review and their own data, organizing the images into five diagnostic classes: healthy, acne, eczema, benign and malignant. Their work used the AlexNet CNN pre-trained for the ImageNet challenge to extract deep convolutional features, which are transformed to error-correcting output codes and then classified by a logistic support vector machine.

Kawahara et al. [12] analyzed and also publicized the 7-point dataset, containing 1,011 images. They developed a multi-modal CNN that receives both clinical and dermoscopic images and also image metadata. Their architecture was based on two pre-trained Inception-v3 CNNs with final $f \times 1 \times 1 \times l$ convolutions that produce the level of abnormality for each criterion in the 7-point checklist (thus inferring the melanoma diagnosis by applying the 7-point score). They classify the images directly in one of five diagnoses: basal cell carcinoma, nevus, melanoma, miscellaneous or seborrheic keratosis. For training, they used a multi-task loss function.

Harangi et al. [13] analyzed the ISIC2018 (also named HAM10000) dataset with 10,015 dermoscopic images.

They designed a CNN architecture based on two pre-trained Inception-v3 CNNs, one for classifying the images in benign or malignant, and another CNN for classification considering seven classes: benign keratosis, dermatofibroma, nevus, actinic keratosis, basal cell carcinoma, melanoma or vascular lesion. Both CNN outputs were merged into a single output layer.

Gessert et al. [14] analyzed the HAM10000 (ISIC2018 training set), 7-point, and ISIC archive (as of 2018) datasets, training with images from HAM only and evaluating with HAM, 7-point, and ISIC images. They subdivided the images into patches, compared the Inception-v3, DenseNet-121, and SE-Resnext50 as baseline deep-feature extractors, and experimented with patch-based Attention, Multi-Crop, and Recurrent output modules. Their work also experimented with four types of dataset balancing, given the increased class imbalance when considering the multi-class classification problem.

Our work differs from previous ones as it achieves an accuracy (Area Under the Curve Receiver-Operating Characteristic (AUC)) at the order of 95% without resorting to general architectures on a transfer-learning basis. Furthermore, we use three different datasets for training and testing, providing heterogeneity with respect to the images' characteristics; this trait is important because the protocol for image acquisition varies from institution to institution, which causes images to be significantly different, despite their common purpose. Inspired by the latest achievements regarding Convolutional Neural Networks, we introduce DermaDL, our own architecture fully trained on skin-lesion images.

Convolutional Neural Networks

In this section, we briefly review techniques Aggregated Transformations and Squeeze-and-Excite, which we use in our design.

Aggregated Transformations

The work of Xie et al. [17] introduced the concept of *Aggregated Transformations*, according to which the tensor is cast for multiple convolutional transformations in the same level of the architecture; activated with non-linearities; and then, aggregated as a tensor with a number of channels. Different from regular convolutional layers, the idea is to slice the tensor into low-dimensional embeddings, and then, fire multiple transformations in parallel. After two convolutions and activations, these slices are concatenated back into an aggregated tensor. This mechanism has not been previously used in the task of skin-lesion classification.

Squeeze-and-Excite

Hu et al. [18] introduce the mechanism of Squeeze-and-Excite, which is capable of improving the expressivity of the convolutional channels throughout a CNN. Its principle is to achieve a global understanding of each channel by squeezing the feature maps into a single numeric value; this is possible with global average pooling, for example. The result is a vector of size C, the number of convolutional channels. This vector passes through a two-layer fully-connected neural network for non-linear activations Relu and Sigmoid, which outputs a vector of the same size. These C values are, then, used to weight the original features of the channels, scaling each channel based on its importance. As we describe next, the combined use of mechanisms Aggregated Transformations with Squeeze-and-Excite rendered a powerful, yet not highly-costly architecture, with a small number of layers if compared to the works presented in Sect. 3.

DermaDL Methodology

We describe the methods used for data preprocessing, followed by the scheme of our neural network design.

Preprocessing

Before training and testing, we proceed with cropping, resizing, and color preprocessing. Since the data are highly unbalanced, we employ data augmentation.

Resizing and Cropping

Skin-lesion datasets contain images with different resolutions, which causes two problems. The resolution many times was prohibitively large, preventing processing in reasonable time, and the variation in size did not allow for a systematic input into a CNN, which demands a constant well-defined resolution. We experimented with different squared resolutions, 384x384, 192x192, 96x96, and 48x48—our experiments demonstrated that 96x96 is sufficient for balancing processing and performance, higher than this resolution costs more processing time without gains; lower than that reduced the performance. Before resizing, we cropped off 15% of the images' periphery, since the lesions are mostly central.

Color Preprocessing

To improve contrast and reduce the effect of noise, we used technique Contrast Limited Adaptive Histogram Equalization (CLAHE) [32], a sort of histogram equalization that considers sections (we used 8x8) of the image to distribute the lightness values; the result corresponds to taking into account the presence of lighter and darker regions separately. The method also uses a contrast limiting scheme—the accumulated histogram value above a certain threshold (we used value 2) is redistributed over all the pixels, preventing noise amplification. The technique improves the local contrast, which favors the detection of edges. Figure 1 illustrates the entire preprocessing for three image samples.
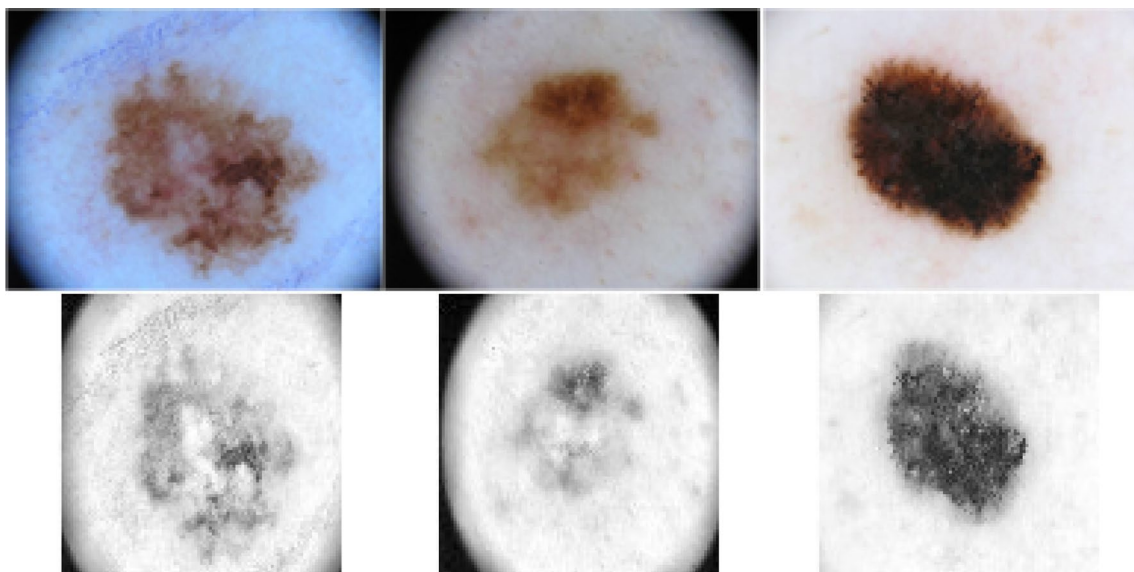
Fig. 1 The images before resizing (upper row), and after resizing with Contrast Limited Adaptive Histogram Equalization (bottom row)

Data Augmentation

All the datasets are highly unbalanced—for example, in the International Skin Imaging Collaboration dataset (refer to Sect. 6.2), the malign cases correspond to only 20% of the images. To cope with that, we augmented the number of malign samples by applying random operations of flipping, brightness, and saturation. We augmented the number of malign images by 5 times, achieving a more even dataset.

Convolutional NN

For the task of skin-lesion classification, we used two techniques that demonstrated to improve the performance of convolutional networks in general: Aggregated Transformations (AG), and Squeeze-and-Excite (SE)—detailed in Sect. 4. After an initial convolution to produce 8 channels, we combine AG and SE into a residual building block—see Fig. 2, which works according to the d_{out} hyper-parameter. Parameter d_{out} states the number of channels of the convolution tensor after the residual block; we used a sequence of three blocks with $d_{out} = \{8, 16, 32\}$. This ensemble demonstrated to achieve highly efficient performances relying on multiple parallel transformation branches, instead of one single very deep network.

We use 8 initial features channels (the *cardinality* hyperparameter); and 8 splits in the Aggregated Transformations block (the *split-out* hyper-parameter). We end up the processing by average pooling each features channel, and then using a dense layer for output, either binary or multi-class.

Fig. 2 Our architecture for skin-lesion classification, which relies on mechanisms Aggregated Transformations (AG), and Squeeze-and-Excite (SE) to define a residual block. We used a sequence of three blocks with channel output dimensionalities $d_{out} = \{8, 16, 32\}$
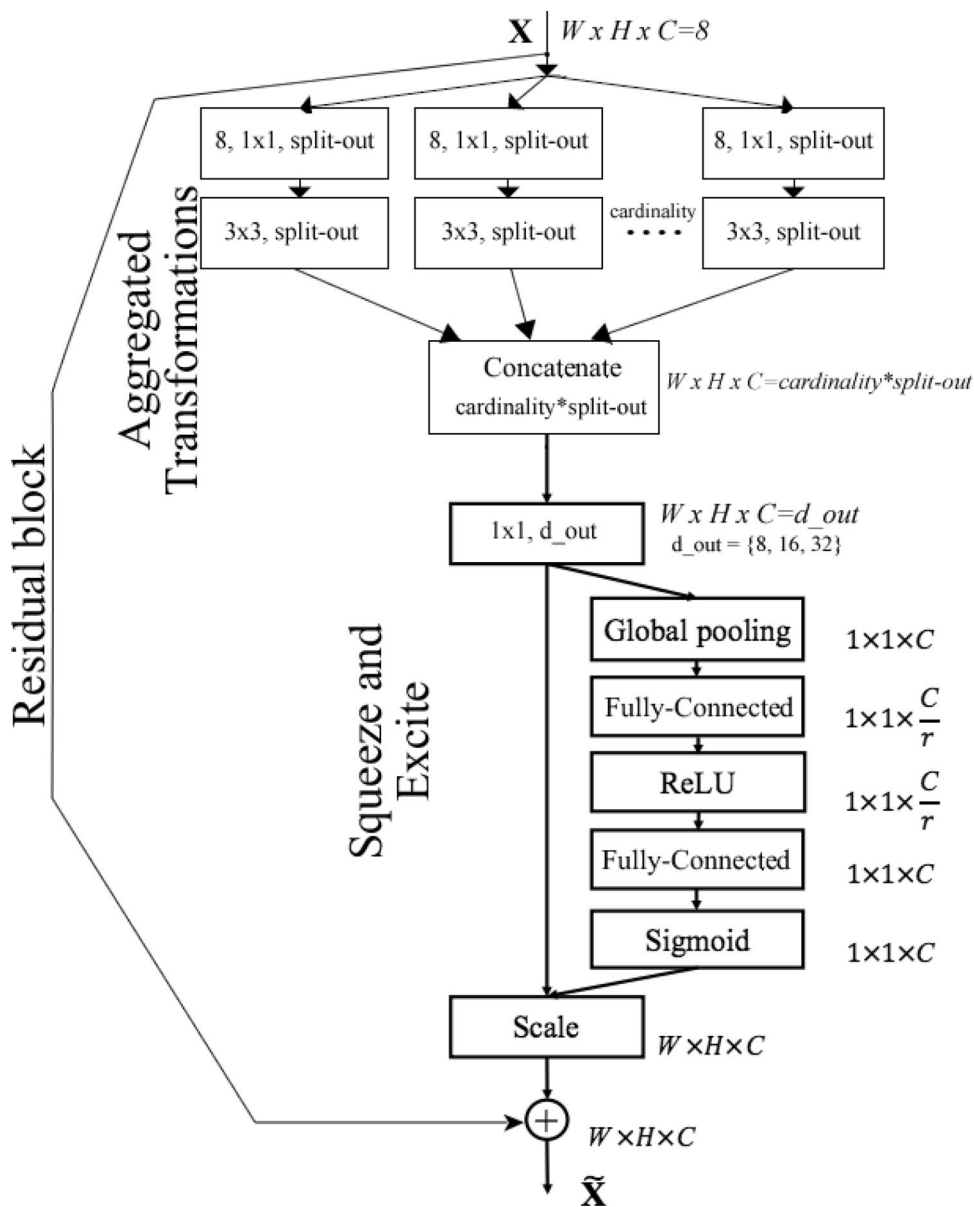
For binary classification, we compute the final probabilities using normalization $softmax(\hat{y}) = \exp(\hat{y})/\sum_i \exp(\hat{y}_i)$. We use loss function $Hinge(\hat{y}_i) = \sum_{i \neq j} \max(0, y_j - \hat{y}_i + 1)$, which suits to binary classification. For multi-class classification, we use the *softmax* output normalization and loss function $H(y, \hat{y}) = -\sum_i [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)]$, which is called cross-entropy and suits to multiple classes. For regularization and convergence, we employ classic L2 normalization with ratio of 0.0005 in the two classifications above. We use batch normalization and Momentum optimization with parameter 0.9. The complete code is available online at https://github.com/jfrjunio/DermaDL.

Binary Melanoma Detection

Implementation

The code was written over framework TensorFlow 1.15 and ran on GPU Nvidia GeForce GTX 1080 Ti; Debian operating system with 8 GB of memory.

Datasets

We use three datasets, ISIC, PH^2, and 7-POINT, achieving an experimental bed with 43,543 images after augmentation. The experiments considered 80% of the data for training, 10% for validation, and 10% for testing.

ISIC

The International Skin Imaging Collaboration (ISIC) endeavor (https://www.isic-archive.com); an international coordinated effort to foster systematic and comparable methods for diagnosing malign skin lesions. The dataset contains images from Melanoma, Seborrheic keratosis, and Nevus. We use ISIC's complete archive with 23.906 images, as of December 2019.

PH^2

This dataset [33] was built for research and benchmarking purposes to facilitate comparative studies on both segmentation and classification algorithms related to dermoscopic images. PH^2 is a dermoscopic image database (https://www.fc.up.pt/addi/project.html) acquired at the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal. It is comprised of 200 images, 20% of which referring to melanoma cases.

7-POINT

This dataset [12] (http://derm.cs.sfu.ca) is a database for evaluating computer-aided prediction methods for skin lesion malignancy. The dataset includes over 2,045 clinical and dermoscopy color images, along with the corresponding structured metadata tailored for training and evaluating computer-aided diagnosis systems.

Experiments

We compare to existing works that heavily relied on architectures tailored for ImageNet and migrated by means of transfer learning to work on the ISIC challenge. As discussed, the ImageNet competition is one of the main experiments to impel the design of architectures to improve the performance on general-purpose object recognition. Due to the efficacy of the best works that appeared on ImageNet, there is a great availability of melanoma detection works based on networks conceived and trained for the ImageNet challenge. Following, we draw a direct comparison to these transfer learned models.

Metrics

Since we deal with a classification problem over a very imbalanced dataset, the metric of simple accuracy is not sufficient—for example, if all the images were classified as negative for melanoma, we would still have high accuracy. Hence, our main metric is the Area Under the Curve Receiver-Operating Characteristic (AUC), which is given by varying a threshold probability score to compute the $Sensitivity = N_{TP}/(N_{TP} + N_{FN}) = $ SE, against $1 - Specificity = 1 - N_{TN}/(N_{TN} + N_{FP}) = 1-$ SP; where N is the number of True/False Positive/Negative classified instances. Once each predicted probability is defined with respect to the answer set, we can draw a confusion matrix and the AUC becomes straight to compute. With this set of metrics, it becomes possible to compare with several works in the literature. Furthermore, due to the strong imbalance of the dataset, it is more critical to correctly identify positive images for melanoma, reducing the False-Negative rate. Therefore, we weighted the loss function to more intensely (60% against 40%) penalize errors related to the positive class.

Results

In Table 1, one can see that, although the previous works demonstrate high AUC, AC, SE, and SP, they perform low for Sensitivity (ratio of true positives). That means that they tend to misclassify a significant number of images that correspond to true melanoma cases, yielding high rates of

Table 1 Comparison to related work [34] considering metrics Area Under the ROC Curve (AUC), accuracy (AC), sensitivity (SE), and specificity (SP) over dataset ISIC

CNN architecture	AUC	AC	SE	SP
AlexNet (ImageNet 2012) [20]	0.859	0.823	0.343	0.969
VGG-16 (ImageNet 2014) [23]	0.892	0.847	0.586	0.865
Inception-v3 (ImageNet 2014) [21]	0.800	0.800	0.355	0.933
ResNet-50 (ImageNet 2015) [24]	0.873	0.723	0.845	0.694
ResNet-101 (ImageNet 2015) [24]	0.869	0.840	0.336	0.986
DermaDL	0.865	0.830	0.905	0.738

Table 2 DermaDL confusion matrix obtained over dataset ISIC

$n = 535$	Conditionpositive	Conditionnegative
Predictedpositive	90.51%(sensitivity)	26.25%(Type I error)
Predictednegative	9.49%(Type II error)	73.75%(specificity)

Table 3 Comparison to related work regarding the network size

CNN architecture	#weights (millions)
AlexNet [20]	61
VGG-16 [23]	16
Inception-v3 [21]	24
ResNet-50 [24]	26
ResNet-101 [24]	250
DermaDL	1.7

"type II" error, possibly the most critical from a clinical perspective.

Contrarily, by observing the confusion matrix depicted in Table 2, one can see that we favored the fine detection of true positives, which slightly reduced our score for specificity (true negatives). Overall, we demonstrated that DermaDL rivals every other technique without resorting to a general-purpose network at the same time that it prioritizes true-positive cases. We emphasize that these numbers are not relative to the ISIC 2017 challenge, but to the complete ISIC archive, after augmentation, partitioned into train, validation, and testing.

In Table 3, one can see that the size of the DermaDL network is pronouncedly smaller than the size of the works based on transfer learning. This result is not surprising; these previous networks were designed to solve a 1,000-classes problem, while DermaDL deals with the binary melanoma/non-melanoma problem. Not only that, we use cutting-edge techniques Aggregated Transformations combined with the mechanism of

Table 4 Experiment considering the entire ISIC archive for training and datasets PH^2 and 7-POINT for testing. For comparison, the first row reproduces the results of Table 1. Training and testing over unrelated datasets revealed unsuccessful

	AUC	AC	SE	SP
ISIC	0.865	0.830	0.905	0.738
PH^2	0.880	0.874	0.450	0.981
7-POINT	0.620	0.635	0.408	0.710

Squeeze-and-Excite organized in a residual block, which confers a superior design in terms of network engineering. These techniques demonstrated a competitive performance with the advantage of using fewer weights (parameters). One of the consequences of our design is that the real-time processing of DermaDL requires a fraction of the computational cost as compared to the other methods; this is because the processing cost is proportional to the number of weights.

Testing on Disjoint Datasets

We also conducted an experiment to evaluate the generalization of our model. We used the ISIC dataset for learning; 90% of which was used for training, and 10% for validation. For testing, we performed experiments on two completely separated datasets, avoiding any kind of biasing—we tested over datasets PH^2 and 7-POINT. We trained for 100 epochs with data shuffling, saving the model with the best validation performance.

The results presented in Table 4 demonstrate that training and testing over dataset ISIC presented a performance better than training over ISIC and testing over datasets PH^2 and 7-POINT, which are completely non-related to the ISIC dataset. By examining a sample of the images of each dataset, we verified that the protocol according to which the images were captured was different for each set of images. The protocols differed to the point of introducing visual particularities that influenced the learning. The images in dataset PH^2 are, almost all, surrounded by a black background originated from some sort of illumination apparatus. Meanwhile, the images in dataset 7-POINT present cases of hair over the lesion, and many times, they have a referential marker close to the lesion, like a grid to allow the inference of the size of the lesion. None of these specific traits is correlated to the malignity of the lesion; they are just part of the image acquisition protocol.

The Need for a Well-Defined Acquisition Protocol

This experiment allowed us to verify that the protocol that guides the image acquisition plays a significant role in

the process of melanoma detection. As one might expect, when the images are added with elements other than the skin lesion, all the images must present the same pattern, be it a marker, a ruler, or an annotation. This is intrinsic to the neural network learning, which is limited with respect to its generalization power; this limitation became flagrant when we trained and tested over three unrelated datasets. It is conclusive that any system designed to aid in melanoma detection must care for providing specific instructions on how to capture the images.

Multi-class Classification

Extended Problem

Skin lesions span to a set of possibilities beyond melanoma; hence, a natural extension of the experimentation over DermaDL is to employ it on the detection of further kinds of lesions. In this section, we evaluate the DL models in a more difficult task by considering multiple diagnoses. We considered the images found in the three experimental datasets, which refer to groups of histological diagnoses: melanoma, carcinoma, keratosis, nevus, and other (see Fig. 3). In these groups, melanoma and carcinoma are malignant, keratosis is either pre-cancerous or benign, nevus and other are benign— the last one corresponds to a group of less frequent lesions. Table 5 illustrates the distribution of each diagnosis in the datasets.

For detecting such lesions, we employed the models listed in Table 9, with multiple outputs in the last layer obtained with normalization softmax and a categorical loss function. The softmax normalization requires the output to be mutually exclusive (multi-class problem), which is intrinsic to our label grouping. We also make use of color to add additional channels of data, improving the potential for a finer prognosis as demanded by the multi-class problem. Accordingly, we used colorized variations of CLAHE.

Table 5 Number of images in each diagnostic class and dataset

| Class	ISIC	7-POINT	PH^2	Total
Melanoma	2169	252	33	2454
Carcinoma	812	42	0	854
Keratosis	1551	45	0	1696
Nevus	18579	575	8	19162
Other	709	97	159	965

Dataset Preparation

For this set of experiments, we relied on a significantly more complex data preparation than that used in the binary problem. We employed three variations of the CLAHE equalization and a class balancing based on color-aware augmentation.

CLAHE Equalization

Similarly to the binary problem, the images in datasets ISIC, PH^2 and 7-POINT were pre-processed, augmented, and split into 80%–10%–10% subsets for training, validation, and testing respectively. The preprocessing starts by erasing the black (or white) borders around the images using median filtering, thresholding in the range [0.1, 0.9], morphological 3x3 opening, finding the bounding rectangle, and cropping the source image. The images are equalized with CLAHE (with parameters threshold=2 and window=8x8) and resized to 144x144. To add color information, we apply CLAHE on the color channels, experimenting on two possibilities beyond the B&W processing, and selecting the one with the best performance. We considered the following three CLAHE approaches, which are illustrated in Fig. 4.

– **B&W**, is the baseline where the image is converted to black and white and CLAHE is applied, thus the color information is discarded;
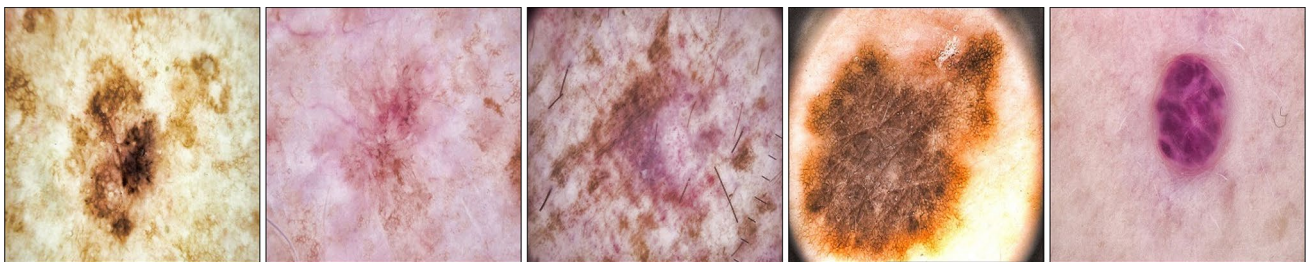
Fig. 3 Examples of the five diagnostic classes in the dataset, respectively, 'melanoma', 'carcinoma', 'keratosis', 'nevus' and 'other' (e.g., angioma)

Fig. 4 CLAHE variations on carcinoma ISIC-0011757: (a) **B&W** increases contrast but erases color information, (b) **RGB** increases the contrast between color channels, making blue–red differences more evident, (c) **Lum** only equalizes the Luminance channel, leaving Hue in the original color scale
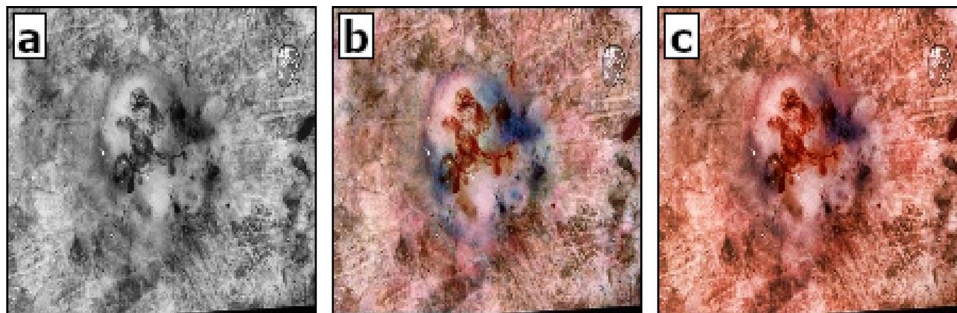
Table 6 Comparison between CLAHE variations considering the multi-class metrics Area Under the ROC Curve (AUC), accuracy (AC), sensitivity (SE), and specificity (SP). The metrics are averaged over all classes

CNN	CLAHE	AUC	AC	SE	SP
DermaDL	B&W	0.747	0.790	0.317	0.908
DermaDL	RGB	0.840	0.819	0.442	0.918
DermaDL	Lum	0.862	0.850	0.473	0.944

– **RGB**, CLAHE is applied in each color channel separately, so different colors are highlighted;
– **Lum**, the image is converted to the HSL color space, then CLAHE is applied only to the Luminance channel; after that, the image is converted back to the RGB color space.

We compared the results of the colorized versions against the B&W baseline on a sample comprising 1/3 of the training data—see Table 6. Approach *Lum* demonstrated the best performance; we used it for the remaining experiments.

Class Balancing via Augmentation

As presented in Table 5, we face a class unbalancing that might hinder the learning process of our algorithms. Hence, we proceeded with an augmentation procedure: for each image, we applied random crops in the range [0.7, 1.0];

width and height shifts with "reflection" fill in the range [-0.1, +0.1] to simulate a slight camera adjustment; vertical or horizontal flip with 50% chance each; brightness in the range [0.8, 1.0]; and saturation changes from the range [-0.1, +0.1] for illumination and tone variation; and a final resizing to 96x96 to fit the network input.

As discussed in Section 6.6, the datasets have quite different acquisition protocols; and are also very unbalanced (e.g., PH^2 has only 8 nevus images), so we balanced the dataset by executing the augmentation procedure to generate 98,150 images. Then, we selected 78,520, 9,815, and 9,815 images for training, validation, and testing. For the multi-class experiment, we used 2.0 million parameters with $c_1 = c_2 = 9$ and $d_{out} = \{58, 116, 232\}$. We refer to this setting as $DermaDL_{multi}$.

Multi-class Metrics

In the multi-class case, the metrics derived from the confusion matrix (Area Under the ROC Curve, accuracy, sensitivity and specificity) are computed separately for each class. We also employ the $SE + SP - 1$ index, which summarizes Sensitivity+Specificity above 50%, i.e., chance. The metrics are then averaged considering all the classes. For example, with k classes, the multi-class AUC is defined as $AUC_{multi} = \frac{1}{k} \sum_{j=1}^{k} AUC(y_j, \hat{y}_j)$. The confusion matrix for the multi-class problem considering setting $DermaDL_{multi}$ is presented in Table 7. The corresponding multi-class metrics appear in Table 8.

Table 7 $DermaDL_{multi}$ confusion matrix obtained over the test set

$n = 9,815$	Conditioncarcinoma	Conditionkeratosis	Conditionmelanoma	Conditionnevus	Conditionother
Predictedcarcinoma	90.98%	13.57%	7.74%	2.23%	6.40%
Predictedkeratosis	3.68%	72.86%	9.94%	4.92%	6.25%
Predictedmelanoma	3.16%	9.73%	71.25%	9.73%	10.11%
Predictednevus	0.62%	2.41%	5.74%	79.84%	6.00%
Predictedother	1.56%	1.43%	5.33%	3.33%	71.24%
	100.00%	100.00%	100.00%	100.00%	100.00%

Table 8 Multi-class metrics per class and source dataset using the confusion matrix of Table 7. Note that PH^2 does not have carcinoma nor keratosis images, and no nevus images into the testing samples

Dataset	Class	AUC	AC	SE	SP	SE+SP-1
ISIC	Carcinoma	0.9780	0.9265	0.9151	0.9295	0.8446
	Keratosis	0.9363	0.8951	0.6899	0.9541	0.6440
	Melanoma	0.9183	0.8814	0.6607	0.9322	0.5929
	Nevus	0.9648	0.9321	0.7922	0.9735	0.7657
	Other	0.9626	0.9280	0.6424	0.9795	0.6219
7-POINT	Carcinoma	0.9571	0.9225	0.7204	0.9632	0.6836
	Keratosis	0.9283	0.8964	0.7416	0.9099	0.6515
	Melanoma	0.9121	0.8423	0.6776	0.9234	0.6010
	Nevus	0.8624	0.9252	0.4860	0.9721	0.4581
	Other	0.9263	0.8423	0.6271	0.9465	0.5736
PH^2	Melanoma	0.9843	0.9521	0.8261	0.9608	0.7869
	Other	0.9610	0.8761	0.8731	0.9167	0.7898
All	Carcinoma	0.9764	0.9285	0.8963	0.9363	0.8327
	Keratosis	0.9341	0.8988	0.6923	0.9501	0.6424
	Melanoma	0.9198	0.8796	0.6658	0.9326	0.5984
	Nevus	0.9581	0.9328	0.7756	0.9733	0.7489
	Other	0.9637	0.9165	0.6784	0.9762	0.6545

Comparison to Related Models

Here, we compare DermaDL to related works that, similarly, deal with the multi-class classification of skin lesions using Deep-Learning models derived from transfer learning. The goal is to demonstrate that DermaDL has the potential to solve the multi-class problem using an architecture designed from scratch, avoiding the drawbacks of transfer learning discussed in Section 1. We notice that our comparison is as precise as possible, but not strictly defined with respect to the other works. Those works vary with respect to the datasets, network inputs (using multiple images or patient metadata), the number of images for training and testing, network and optimizer meta-parameters, evaluation metrics, and diagnostic categories. The common aspect is that of the nature of the images, learning task, and the Deep-Learning techniques, which confer a valid, but not absolute, comparative perspective. As future work, we envision the comparison of multiple architectures over exactly the same settings.

To simplify our comparison, we report the multi-class metrics in Table 9, as made available by the original authors.

In our experiments with DermaDL, presented in the last two rows of the table, we present evidence that our architecture achieves metric results with similar magnitudes. For *DermaDL_multi*, metric AUC achieved the highest reported mark, and index SE+SP-1 achieved the third-highest mark. But notice that, as detailed in Section 7.7, this performance was achieved with an architecture one order of magnitude smaller than that of previous works. In other words, DermaDL demands a fraction of the processing required by former works, with much room for improvement.

Discussion

Broadening the goal to a multi-class classification significantly increases the problem complexity, which allowed us to draw interesting conclusions. We observed that a more sophisticated preprocessing based on colorized CLAHE was able to improve the AUC and the Sensitivity by significant amounts considering metrics AUC, AC, SE, and SP, as presented in Table 6. We also verified significant results in comparison to much larger transfer-learning-based architectures

Table 9 Comparison to related work on the multi-class classification of skin lesions, considering the metrics Area Under the ROC Curve (AUC), accuracy (AC), sensitivity (SE), specificity (SP), and index SE+SP-1. The metrics are averaged over all classes

CNN	Classes	#Images	AUC	AC	SE	SP	SE+SP-1
AlexNet [11]	5	9,144	N/A	0.945	0.840	0.964	0.804
Inception-v3 [12]	5	1,011	0.896	N/A	0.604	0.910	0.514
Inception-v3 [13]	7	10,015	0.923	0.946	0.639	0.957	0.596
Inception-v3 [14]	7	10,015	N/A	N/A	0.757	0.960	0.717
DermaDL_multi	5	98,150	0.950	0.911	0.742	0.954	0.695

Table 10 Size of the network architectures in the multi-class problem

Architecture	Input size	#Weights ($\times 10^6$)
AlexNet [11]	224x224	62.38
Inception-v3 [12–14]	299x299	23.85
DermaDL$_{multi}$	96x96	2.00

[11–14] using more advanced architectural techniques. As one can see in Table 10, DermaDL is one order of magnitude smaller than previous works, indicating that it is possible to train, from scratch, a small network (2 million parameters with 96x96 input size) and still achieve a good overall performance in the multi-class scenario (0.950 AUC for *DermaDL$_{multi}$*). For reference, the smallest state-of-the-art networks in the ImageNet dataset, including the architectures engineered by AutoML methods, have, at least, 5 million parameters [35].

Conclusions

We presented a Convolutional Neural Network that innovates using mechanisms Aggregated Transformation and Squeeze-and-Excite combined into a residual processing block. Our architecture achieved results comparable to the state-of-the-art without transferring the learning from existing general-purpose architectures, and at the same time, demonstrating an ensemble with far fewer layers and parameters than previous works. Instead, we designed a processing flow from scratch, exclusively training it with skin-lesion images. The complete code is available at https://github.com/jfrjunio/DermaDL.

We experimented with datasets ISIC, PH^2, and 7-POINT demonstrating that the use of cutting-edge Convolutional Neural Network techniques can rival previous results using just a fraction of the number of weights. Our architecture was experimented for melanoma detection (a binary problem), and skin-lesion classification (a multi-class problem), achieving results comparable to the state-of-the-art. We also demonstrated the importance of the image acquisition protocol by training and testing with unrelated datasets. Our work allows for future works to fine-tune a network specifically tailored for skin-lesion classification, progressively improving it similarly to what was done for general-purpose object recognition, year after year, in the ImageNet competition.

Declarations

Conflict of interest statement On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

1. Cheng YI, Swamisai R, Umbaugh SE, Moss RH, Stoecker WV, Teegala S, Srinivasan SK. Skin lesion classification using relative color features. Skin Res Technol. 2008;14(1):53–64.
2. Menzies SW, Bischof L, Talbot H, Gutenev A, Avramidis M, Wong L, Lo SK, Mackellar G, Skladnev V, McCarthy W, et al. The performance of solarscan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. Arch Dermatol. 2005;141(11):1388–96.
3. T. A. C. Society, Key statistics for melanoma skin cancer. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html. Accessed Dec 2020.
4. Schadendorf D, van Akkooi ACJ, Berking C, Griewank KG, Gutzmer R, Hauschild A, Stang A, Roesch A, Ugurel S. Melanoma. The Lancet. 2018;392(10151):971–84.
5. Jerant AF, Johnson JT, Sheridan CD, Caffrey TJ. Early detection and treatment of skin cancer. Am Fam Phys. 2000;62(2):357–68.
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
7. Codella NC, Nguyen Q-B, Pankanti S, Gutman DA, Helba B, Halpern AC, Smith JR. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J Res Dev. 2017;61(4/5):5–1.
8. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, Watanabe R, Okiyama N, Ohara K, Fujimoto M. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol. 2018;180(2):373–81.
9. Menegola A, Tavares J, Fornaciali M, Li LT, de Avila SEF, Valle E. RECOD titans at ISIC challenge 2017. arXiv:1703.04819
10. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol. 2018;138(7):1529–38.
11. Hameed N, Shabut AM, Hossain MA. Multi-class skin diseases classification using deep convolutional neural network and support vector machine. In 2018 12th International Conference on Software, Knowledge, Information Management Applications (SKIMA), 2018, pp. 1–7.
12. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inf. 2018; 23(2):538–546.
13. Harangi B, Baran A, Hajdu A. Assisted deep learning framework for multi-class skin lesion classification considering a binary classification support. Biomed Signal Process Control, 2020; v. 62, p. 102041.
14. Gessert N, Sentker T, Madesta F, Schmitz R, Kniep H, Baltruschat I, Werner R, Schlaefer A. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. IEEE Trans Biomed Eng., 2019; v. 67, n. 2, p. 495–503.

15. Torrey L, Shavlik J. Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global. 2010;242–264.

16. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg A, Fei-Fei L. ImageNet large scale visual recognition challenge. Int J Comput Vis (IJCV). 2015;115(3):211–52.

17. Xie S, Girshick R, Dollar P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: IEEE conference on computer vision and pattern recognition, 2017;pp. 1492–1500.

18. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018;pp. 7132–7141.

19. Abbasi, Naheed R., et al. Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. JAMA 2004; 292(22), 2771–2776.

20. Krizhevsky A, Sutskever I, Hinton G. "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems. Curran Associates. 2012;1097–105.

21. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. "Rethinking the inception architecture for computer vision," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp. 2818–2826.

22. Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra N, Kittler H, Halpern A. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging. In IEEE Biomedical Imaging, 2018; pp. 168–172.

23. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556

24. 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, 2016;pp. 770–778

25. Zakhem GA, Motosko CC, Ho RS. How should artificial intelligence screen for skin cancer and deliver diagnostic predictions to patients? JAMA Dermatol. 2018;154(12):1383–4.

26. Ruiz D, Berenguer V, Soriano A, Sanchez B. A decision support system for the diagnosis of melanoma: a comparative approach. Expert Syst Appl. 2011;38(12), 15217–15223

27. Yu L, Chen H, Dou Q, Qin J, Heng PA. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging. 2017;36(4):994–1004.

28. Ganzeli H, Bottesini J, Paz L, Ribeiro M. Skan: Skin scanner - system for skin cancer detection using adaptive techniques. IEEE Lat Am Trans. 2011;9(2):206–12.

29. Aswin RB, Jaleel JA, Salim S. Hybrid genetic algorithm - artificial neural network classifier for skin cancer detection. In Control instrumentation: communication and computational technologies, 2014; 1304–1309.

30. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Jafari MH, Ward K, Najarian K. Melanoma detection by analysis of clinical images using convolutional neural network. In IEEE Engineering in Medicine and Biology Society, 2016;pp. 1373–1376.

31. Majtner T, Yildirim-Yayilgan S, Hardeberg JY. Combining deep learning and hand-crafted features for skin lesion classification. In: International Conference on Image Processing Theory, Tools and Applications, 2016; pp. 1–6.

32. Gonzalez R, Woods R, Eddins S. Digital image processing using MATLAB. Pearson, 2004.

33. Barata C, Ruela M, Francisco M, Mendonca T, Marques J. Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE systems Journal, 2013; v. 8, n. 3, p. 965–979.

34. Li Y, Shen L. Skin lesion analysis towards melanoma detection using deep learning network. Sensors. 2018;18(2):556.

35. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019; p. 6105–6114.

CARDIAC MRI ANALYSIS

In this chapter, we reproduce the following article:

Lima *et al.* (2022). *Full Motion Focus: Convolutional Module for Improved Left Ventricle Segmentation over 4D MRI*. Lecture Notes in Computer Science, Springer. Presented in ICIAP 2022, Leece, Italy.

In this article we described Full-Motion-Focus (FMF), a module that detects the heart motion in the 4D Magnetic Resonance Imaging (MRI) sequence, and highlights a Region of Interest (RoI) by focusing a Radial Basis Function (RBF) on the estimated motion field. MRI is a widely known medical imaging technique used to assess the heart function.

MRI applies a strong magnetic field (1.5–3 T) to the whole body and detects changes in the magnetization field of the water molecules in the observed region, after applying controlled radiofrequency pulses. The field change is acquired and processed by inverse Fourier Transform, which returns a tri-dimensional image of soft tissues in common Euclidean coordinates. The process is repeated several times to collect data in all regions of the heart.

By adding the time dimension, MRI is able to capture a full heartbeat, so called 4D or cine MRI. The pictures are usually captured in 5-10 mm thick slices spaced by 5-10 mm gaps, in intervals of 20–100 ms between frames. The acquisition head can be positioned to capture full resolution in short or long axis. This method can take high-resolution pictures of the heart without any damage, and is considered the gold standard in medical literature.

Among the many computer methods that process MRI images as reviewed by Peng *et al.* (2016), Deep learning (DL) models can also perform several tasks in Cardiac MRI (CMR) images with good efficacy, such as segmentation, estimation, and detection of diseases. The most basic task is tissue segmentation, which is used to measure muscle and cavity dimensions, estimate muscle mass and dislocation, then assert the heart function in terms of how much blood is pumped and if the ventricle walls have normal movement.

For example, heart assessment guidelines, e.g. reported in Montera *et al.* (2022), use the Left Ventricle Ejection Fraction (LVEF) as a central measure of the heart function. LVEF below 50% is interpreted as an impairment, and below 35-45% it becomes a risk factor for diseases such as I50–"*Heart Failure*" (World Health Organization, 2016).

This measure is estimated from the left ventricle volumes measured on a segmented heart image—LVEF is the percent ratio of stroke volume (SV) and end-diastolic volume (EDV), taken in the two extreme instants of the heartbeat (systole and diastole). Foley *et al.* (2012) remarks that the segmentation is a critical step, and the final LVEF estimation is sensitive to the imaging technique and precision of the algorithm, motivating DL research.

Many DL models are based on convolutional neural networks (CNN), and were improved by detecting a RoI either automatically or by hand. The automatic methods either add pre or post processing, or can be embedded in the network architecture itself. However, the embedded approaches such as multi-level modules or attention mechanisms—e.g. those employed in Inception from Szegedy *et al.* (2015)—require extensive network training and data volumes.

In the paper we proposed FMF, an algorithm that delineates a spherical RoI with a RBF around the heart motion in 4D MRI, whose boundaries can be adjusted by thresholding a probability function. The cutoff threshold was defined with a heuristic based on the image statistics. We took care to define the algorithm solely with convolution operations so it is possible to embed the method inside a convolutional neural network.

We experimented and evaluated FMF for left ventricle segmentation on three CMR datasets, observing that the proposed RoIs covered 99.7% of data labels (Recall score), improved the CNN segmentation by 1.7 (mean Dice score, χ^2 p<0.001) after the RoI extraction, and improved the overall training speed by 2.5 times (+150%). This was an incremental result that demonstrates a Deep learning application in another medical imaging domain.

Full Motion Focus: Convolutional Module for Improved Left Ventricle Segmentation Over 4D MRI

Daniel M. Lima[1,2(✉)] , Catharine V. Graves[2] , Marco A. Gutierrez[2] ,
Bruno Brandoli[3] , and Jose F. Rodrigues Jr.[1]

[1] Institute of Mathematics and Computer Science, Universidade de Sao Paulo,
Sao Carlos, SP, Brazil
danielm@usp.br
[2] Laboratorio de Informatica Biomedica, Instituto do Coracao, Hospital das Clinicas,
Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, SP, Brazil
[3] Dalhousie University, B3H 4R2 Halifax, NS, Canada

Abstract. Magnetic Resonance Imaging (MRI) is a widely known medical imaging technique used to assess the heart function. Over Cardiac MRI (CMR) images, Deep Learning (DL) models perform several tasks with good efficacy, such as segmentation, estimation, and detection of diseases. Such models can produce even better results when their input is a Region of Interest (RoI), that is, a segment of the image with more analytical potential for diagnosis. Accordingly, we describe *Full Motion Focus* (*FMF*), an image processing technique sensitive to the heart motion in a 4D MRI sequence (video) whose principle is to combine static and dynamic image features with a Radial Basis Function (RBF) to highlight the RoI found in the motion field. We experimented *FMF* with the U-Net convolutional DL architecture over three CMR datasets in the task of Left Ventricle segmentation; we achieved a rate of detection (Recall score) of 99.7% concerning the RoIs, improved the U-Net segmentation (mean Dice score) by 1.7 ($p < .001$), and improved the overall training speed by 2.5 times (+150%).

Keywords: Cardiac MRI · Motion · Deep learning · Localization · Segmentation

1 Introduction

Magnetic resonance imaging (MRI) is a medical imaging technique used to capture volumetric image sequences of internal soft tissues, such as cardiac muscles. In comparison to X-ray imaging (XR) and computed tomography (CT), MRI provides images with improved structural details via finer spatial resolutions. Cardiac MRI (CMR) focuses on the heart, allowing trained cardiologists to measure heart parameters, for example, the mass of the cardiac muscle (myocardium mass), volumes of blood cavities (atrial and ventricular volumes) and volume of

blood pumped per heartbeat (ejection fraction) [9]. Those parameters are used to assess how healthy is the heart, by recognizing early conditions and signs before the onset of infarcts and other complications.

Due to the complexity of CMR images, comparably complex techniques are required to produce detailed analyses. One of these techniques is deep learning (DL). Many of the tasks related to the cardiac function analysis have benefited from DL methods—for example segmentation of structures [1], estimation of heart parameters [17], and detection of diseases [7]. For even better results, research in DL has pointed out that models based on convolutional neural networks (CNN) had higher efficacy when provided with regions of interest (RoI) either explicitly or implicitly [17]. The RoI proposal is a preprocessing step whose goal is to identify the most prominent regions of an image for discovering clinically relevant artifacts.

The explicit RoI proposal approaches usually follow a combination of methods, for example: (a) pipelining a segmentation and a regression network; (b) preprocessing the input with a region proposal algorithm [5] or with a CNN [16]; or (c) by using manual cropping [18]. The implicit RoI detectors are added to the DL network abstracted as additional operators and variables; e.g. multiscale Inception [14] and attention [15] modules, which benefit from the RoI to down-weight less-informative neurons and inputs inside the network. Inception modules weight convolutions of different sizes, while attention modules assign a weight to each feature channel. This additional neural information processing guides which input features or channels shall have more weight.

In this paper, we develop a module that highlights regions in the image sequence by analyzing the motion field using a radial basis function (RBF) [8]. In our experiments, we analyze our method by using the RBF for cropping the input before having it processed by a pretrained segmentation U-Net convolutional network [10]. Our methodology is an innovation in the task of region proposal for CMR analysis; we demonstrate results that justify the use and further investigation of the employed principles. We named it after its working mechanism as Full Motion Focus (*FMF*).

2   Theory and Related Work

2.1   Cardiac MRI

MRI is the most precise medical imaging technique for examination of the heart structures, it records heart images along a complete heartbeat cycle [2,12]. In practice, the magnetization signal is triggered by a reference pulse, then captured several times for noise reduction and, finally, reconstructed by inverse Fast Fourier Transform. The resulting image is usually visualized in slices along a positional axis: the long axis has a frontal or lateral view of the heart, and the short axis aligns to a cross-sectional plane.

The short-axis view is split in three regions: the base or basal region near the top where blood vessels connect to the heart (slice B); the middle or medial in between (slice M); and the apex or apical region at the bottom tip of the heart

(slice A)—refer to Fig. 1. The normal human heart has four chambers: right atrium (RA), right ventricle (RV), left atrium (LA), and left ventricle (LV). The atria receive blood and the ventricles pump it out of the heart. Even though all chambers are important, the LV is of special interest because it is the cardiac muscle that does the "heavy lifting" of pumping oxygenated blood from the lungs to the whole body. In the short-axis view, the LV appears as a ring shape, whose thickness and internal volume measurements are essential to estimate the myocardium mass and ejection fraction, respectively.
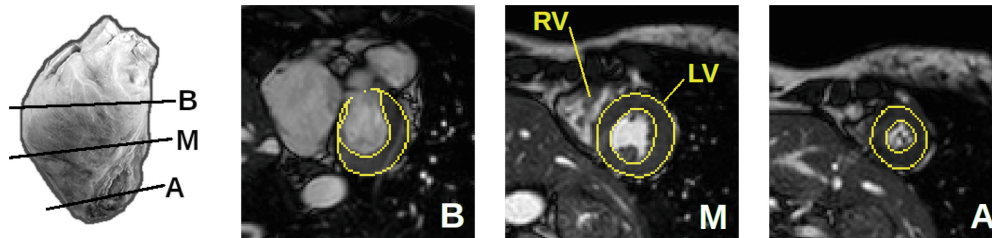
Fig. 1. Picture of heart, and examples of CMR cross-sections in the Base (B), Middle (M) and Apex (A) regions, with LV boundaries. Adapted from Saber *et al.* [11].

2.2   Computer Methods for CMR Analysis and LV Segmentation

Computer methods for functional analysis of CMR images, as reviewed by Peng *et al.* [9], were organized in three ways: image-driven, model-driven, and by direct estimation. Further subdivisions of the LV segmentation spans five groups: (1) image processing methods such as thresholding, morphology operators, and region-growing; (2) pixel/voxel-based classification by Gaussian mixture models, neural networks, k-means, k-NN, or SVM; (3) active contours (snakes), deformable models, level sets, and motion tracking; (4) PCA or ICA with strong priors from anatomic heart models; and (5) direct estimation by, e.g., latent discriminant analysis combined with SVM. This work refers to category 2 as it employs a pixel-based classification by a neural network.

Recent approaches for LV segmentation use CNNs (such as U-Net) experimented over many datasets and methodology combinations. U-Net [10] is a general segmentation model which combines a tower of downscaled-then-upscaled deep representations. For CMR images, U-Net displayed good results (89% Dice) when trained from min-cut priors [4]. A U-Net architecture with residual blocks and optical flow information achieved 89%, 95% and 85% Dice in the base, middle, and apex regions of the heart, respectively, in the work of Yan *et al.* [19]. In the work of Wu *et al.* [16], the authors combine a custom CNN for region proposal with a U-Net segmentation to achieve 95% Dice. Overall, the combination of region proposal to U-Net had good results in particular datasets, but still has room for improvement when the evaluation generalizes across multiple datasets. Different from former works, our methodology, *FMF*, uses RBFs to propose RoIs that will aid a CNN processing in the task of LV segmentation, which we demonstrate with experimental U-net improvements in multiple datasets.

3   Materials and Methods

FMF starts with a 4D image input $\mathbf{x} = I(t, x, y, z)$, that is, a sequence of volumes (frames) each one with a time t coordinate. Initially, we normalize $\mathbf{x}$ to produce sequence $\mathbf{x}^*$ with frames in a format more adequate for Neural Network processing—see Sect. 3.1. From $\mathbf{x}^*$, we extract static visual features to produce $\mathbf{x}_s$, and motion estimation to produce $\mathbf{x}_t$, detailed in Sect. 3.2. Next, we apply two sets of weights: w_s are weights related to visual features, and w_t are weights related to the motion (time). After weighting, both features are combined in tensor $\mathbf{v}$. Then, as presented in Sect. 3.3, we compute the center voxel μ_v defined by the weighted sum of all the voxels' coordinates. The segmentation map $\mathbf{y_S}$ is produced by applying a threshold to $\mathbf{v}$ and extracting the bounds of non-zero voxels. Afterwards, we compute the scale σ_v given by the standard deviation of all the voxels' distances from center μ_v. At this point, refer to Sect. 3.4, we can apply a radial basis function at center μ_v with radius σ_v, computing $\mathbf{y_L}$, then we scale the region defined by $\mathbf{y_L}$ to the CNN input shape. In Sect. 3.6, we explain the neural network processing, its parameters, and training.

3.1   Intensity Normalization

We normalize the image intensities between 0 and 1 by subtracting the minimum value and dividing by the range of values, using an small constant ϵ to avoid division-by-zero:

$$\mathbf{x}^* = \frac{\mathbf{x} - \min(\mathbf{x}) + \epsilon}{\max(\mathbf{x}) - \min(\mathbf{x}) + \epsilon} \tag{1}$$

3.2   Visual Features and Motion Estimation

For visual features extraction, we consider the statistical mean and standard deviation obtained with $3 \times 3 \times 3$ kernels. The mean image I_μ is computed using the convolution operation with mean kernel M, defined as:

$$M = \frac{1}{1 \cdot 3 \cdot 3 \cdot 3} \cdot J_{1,3,3,3} \tag{2}$$

where $J_{1,3,3,3}$ is a $1 \times 3 \times 3 \times 3$ matrix-of-ones. That is, M is just the arithmetic mean of a 3^3 matrix, in a convolution form. With kernel M, we compute I_μ as:

$$I_\mu = I * M \tag{3}$$

In turn, the standard deviation image I_σ is computed by taking the differences between image I and mean image I_μ, then squaring the differences element-wise (Hadamard power); after that, we apply a convolution operation with the mean kernel M, before taking the Hadamard square-root, as follows:

$$I_\sigma = [(I - I_\mu)^{(2)} * M]^{(1/2)} \tag{4}$$

Accordingly, the mean image I_μ and the standard deviation image I_σ define $\mathbf{x}_s = \{I_\mu, I_\sigma\}$. Notice that we express the Hadamard powers using the definitions and notations defined in the work of Fallat and Johnson [3].

Motion is estimated by function $E(I)$—in Eq. 5, given by the root-mean-squared differences of intensity along the time coordinate, where T is the time interval (or number of frames) in image I, and S_t is the Sobel kernel w.r.t. time, instead of the default S_x and S_y spatial Sobel kernels. This function is related to the magnitude of the optical flow vector field [6] in each voxel, as follows:

$$\mathbf{x}_t = E(I) = \sqrt{\frac{1}{T}\int_{t=0}^{T}\left(\frac{\partial I}{\partial t}\right)^2 dt} \approx \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}[I(t)*S_t]^{(2)}} \tag{5}$$

Figure 2 illustrates a sequence of six CMR frames spaced by 1/6 of the cardiac cycle (upper row in the figure), and the respective absolute derivatives in each point along the time dimension (lower row) as computed by Eq. 5. Figure 3a shows the total motion estimate in the sequence, while Fig. 3b presents the cumulative energy histogram. By combining both static and motion features $\mathbf{x}_s$ and $\mathbf{x}_t$, we obtain:

$$\mathbf{v} = w_s\mathbf{x}_s + w_t\mathbf{x}_t \tag{6}$$

The weights w_s and w_t are initialized to 0.1 and 0.9 respectively for $\mathbf{x}_s$ and $\mathbf{x}_t$, i.e., although we emphasize the motion features, we also include the static visual features, which will address the problematic cases when the heart has limited motility, which might be the case for heart complications.
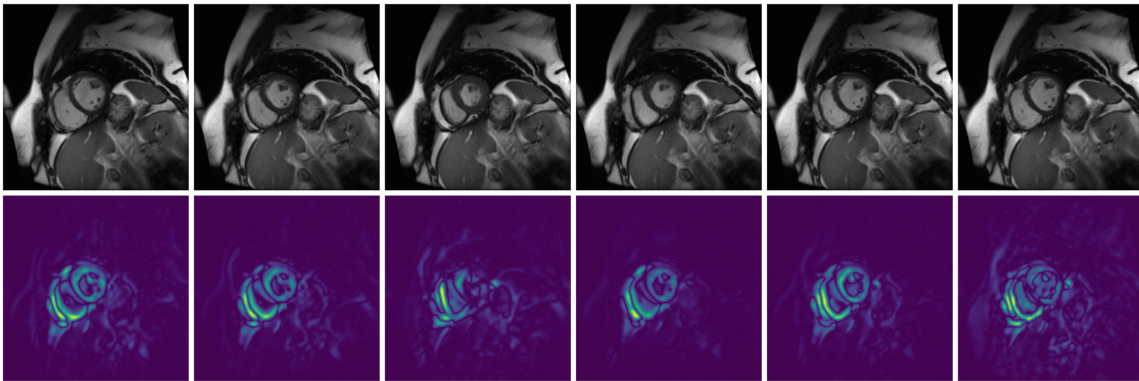
Fig. 2. Images in the medial slice, showing six frames (upper) and absolute time derivatives (lower). This figure only shows medial slices, but *FMF* is defined for volumes.

3.3 Center and Scale Computation

The center of energy μ_v, formalized by Eq. 7, is the voxel defined by the energy-weighted sum of all the voxels' coordinates $\mathbf{r}$ in dimensions x, y, z; that is, every
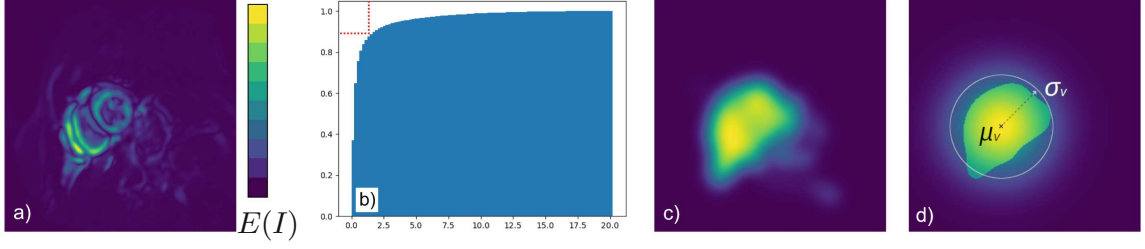
Fig. 3. a) Motion estimate $E(I)$ in whole cardiac cycle. b) Cumulative histogram's shoulder lies near threshold $p = 0.9$. c) $E(I) > 90\%$ of Figure a, convolved with Gaussian($\sigma = 5$). d) The RBF is fitted to this region with center μ_v and radius $\propto \sigma_v$.

voxel i has $\mathbf{r}_i = \langle x_i, y_i, z_i \rangle \mid 0 \le x_i < \text{width}, 0 \le y_i < \text{height}, 0 \le z_i < \text{slices}, 1 \le i \le N$, for N is the number of voxels in the image, as follows:

$$\mu_v = \langle \bar{x}, \bar{y}, \bar{z} \rangle = \frac{\sum_{i=1}^{N} \mathbf{v}_i \cdot \mathbf{r}_i}{\sum \mathbf{v}} \tag{7}$$

The segmentation $\mathbf{y_S}$ is given by the thresholded image from the previous step, see Fig. 3c. The scale estimate σ_v is the cube root of the volume (in voxels) of the thresholded image considering the values above the 90% percentile:

$$(\mathbf{y_S})_i = \mathbb{1}[\mathbf{v}_i > Q(0.9, \mathbf{v})] \tag{8}$$

$$\sigma_v = \frac{3}{\|\mathbf{r}_{\max}\|} \sqrt[3]{\sum_{i=1}^{N} (\mathbf{y_S})_i} \tag{9}$$

where the indicator function $\mathbb{1}[c]$ returns 1 if c is true, or 0 otherwise; and $Q(p, v)$ is the quantile function, returning the maximum of the lowest p (%) values in v.

3.4 Segmentation and Localization Focus

The segmentation focus $\mathbf{y_S}$ is derived from thresholded $\mathbf{v}$ (Eq. 8), and the localization $\mathbf{y_L}$ is found by fitting an RBF to $\mathbf{y_S}$. The radius d_i is the distance from the center μ_v to each voxel i. The Euclidean distance (L_2-norm $\|\cdot\|$) was used:

$$d_i = \left\| \frac{\mathbf{r}_i - \mu_v}{\mathbf{r}_{\max}} \right\| = \sqrt{\left(\frac{x_i - \bar{x}}{x_{\max}} \right)^2 + \left(\frac{y_i - \bar{y}}{y_{\max}} \right)^2 + \left(\frac{z_i - \bar{z}}{z_{\max}} \right)^2} \tag{10}$$

The chosen RBF is a Gaussian ϕ of the voxels' distances:

$$(\mathbf{y_L})_i = \phi(i) = \exp\left[-(d_i/\sigma_v)^2 \right] \tag{11}$$

Both outputs $\mathbf{y_S}$ and $\mathbf{y_L}$ are illustrated in Fig. 3d. According to the framework explained so far, we can design a focus for many different objects in the images by changing the functions for image feature extraction, motion estimation, center, scale, and RBF. It is also possible to detect multiple objects or objects of complex shapes by a mixture of RBF models.

3.5   Crop and Scaling

This step performs the final preparation of the image so that it will fit the CNN input shape; this is necessary because the MRI images may have diverse dimensions. Besides, the RBF estimation is executed in the original resolution, so the proposed region must be cropped then adjusted, followed by an intensity normalization. For CNNs with fixed input dimensions, we rescale the images using bicubic spline interpolation. For CNNs with variable input shape, we only adjust the image proportions as requested by the model, e.g. the U-Net we use has five max-pooling layers with a down-scaling factor of 2, which means the input dimensions should be multiples of 2^5.

Table 1. Overview of the CMR datasets. n = number of patients; Sxy = spatial resolution (pixel spacing) in the axial plane (mm/pixel); Sz = slice resolution (mm/pixel).

Dataset	n	width	height	slices	frames	Sxy	Sz
LVSC	100	138–512	138–512	8–24	18–35	0.68–2.14	6–10
ACDC	100	154–428	154–512	6–18	12–35	0.70–1.92	5–10
M&Ms	320	196–548	192–512	6–20	18–36	0.68–1.82	5–10

3.6   Segmentation CNN

In this section, we employ method *FMF* in combination with the 2D U-Net CNN [10], a consolidated technique widely tested for image processing—refer to Sect. 2.2; the methodology, though, is suitable to any other CNN. The *FMF* pipeline executes along the whole cardiac volume and, after cropping, the RoI pass to the segmentation CNN, which in this case is the 2D U-Net. The CNN was trained with outputs obtained with method *FMF* during 30 epochs using an adaptive momentum optimizer, initial learning rate $\eta = 0.001$, Nesterov $\beta_1 = 0.9$, L_∞-decay $\beta_2 = 0.999$, and loss function binary cross-entropy plus the Sørensen-Dice coefficient (DSC):

$$\mathcal{L}(y, p) = -[y \cdot \log(p) + (1-y) \cdot \log(1-p)] + DSC(y, p) \tag{12}$$

where DSC is a performance metric defined in Sect. 4.1, Eq. 14 .

3.7   Datasets

To evaluate method *FMF*, we used three CMR datasets specified across the short-axis orientation. The metadata in the datasets include: a) binary masks for the LV and RV; b) physiological parameters such as myocardium mass, ventricle area, volume, ejection fraction, thickness, and dimensions of structures; c) image acquisition parameters such as spatial resolution (mm), temporal resolution (frames per cardiac cycle) and slice gaps (mm). Not all of the datasets encompass the same information; following, we provide more details, with a summary presented in Table 1.

- **LVSC** - Cardiac Atlas Project (CAP) 2011 LV Segmentation Challenge [13]; we used the 100 patients in the training set with LV masks on all frames;
- **ACDC** - MICCAI 2017 Automated Cardiac Diagnosis Challenge [1]; it was created from clinical data, including sequences of 100 patients with RV and LV masks in two frames, from the University Hospital of Dijon (France) over 6 years;
- **M&Ms** - MICCAI 2020 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge [11]; this database was collected from six hospitals in Spain, Canada, and Germany using several MRI scanners (Siemens, GE, Philips, and Canon), it includes RV and LV masks in two frames—we used 320 patients for which the labels are publicly available.

4 Experiments and Results

We evaluated our methodology by analyzing the selected datasets and comparing the results of two CNNs: one base U-Net CNN on the raw images without the *FMF* RoI proposal, and another U-Net CNN on the images processed by *FMF*. After running *FMF* and cropping the RoI, the cropped frames were passed to the U-Net. The myocardium labels were obtained by subtracting epicardium and endocardium masks. The datasets were individually split with 75% for training and 25% for validation of models; for testing, we performed an all-versus-all scheme, that is, fitting the CNNs to the training set of one dataset and testing on the validation set of another dataset.

The experiments ran in a computer with Intel i7-7700k CPU and NVIDIA Titan-Xp GPU. Our software was written in Python, scipy, tensorflow, and matplotlib. The steps were: load an image; compute features; fit RBF; crop and resize both the input image and the label mask; Base CNN prediction on the original image and *FMF*-CNN prediction on the cropped image; then, evaluate the predicted masks against the labels using the following metrics.

4.1 Metrics

We compared three indices: (1) Recall, the proportion of the labels that was preserved in the *FMF* region proposals; it is also known as Sensitivity or True Positive Rate—TPR as defined by Eq. 13, which aims to verify if the bounding boxes cover the labels entirely; (2) the Sørensen-Dice coefficient—DSC as defined by Eq. 14, which is equivalent to the F1-score (average of Precision and Recall) that refers to the segmentation output when not-using vs using method *FMF*; (3) speedup, the ratio of time taken by the CNN when not-using vs using method *FMF*—it is defined as $(t_{\text{base}}/t_{\text{ours}})$.

$$TPR(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y|} = \frac{TP}{TP + FN} \tag{13}$$

$$DSC(y, \hat{y}) = \frac{2|y \cap \hat{y}|}{|y| + |\hat{y}|} = \frac{2TP}{2TP + FP + FN} \tag{14}$$

Recall ranges from 0, when none of the marked voxels are detected; to 1 when 100% of the marked voxels are detected. DSC ranges from 0, for no intersection; to 1 for a perfect match between y and $\hat{y}$. Speedup is a positive ratio with value =1 when times are equal; >1 when *FMF* is faster; and <1 when *FMF* is slower.

4.2   Results

Table 2 presents the results of the *FMF* RoI proposal as Recall, and the results of the *FMF*-CNN segmentation as Dice score. The Recall, that is, the ability to identify the regions of interest, was nearly perfect for all the datasets. Concerning the Dice score, the U-Net trained on dataset M&Ms (the largest dataset) was significantly improved when using the *FMF* RoI proposal, with a mean performance increase of +7.2 (percent points) considering all the datasets.

Table 2. Performance results. In the first (left-most) column, the name of the dataset and the Recall metric. In the remaining columns, the DSC (Dice score) for the Base CNN segmentation and for the segmentation obtained after the *FMF* RoI proposal.

Train	Test	Base	FMF	ΔDSC	Train	Test	Base	FMF	ΔDSC
M&Ms 99.75% recall	LVSC	62.4%	70.9%	+8.5		LVSC	50.1%	44.8%	−5.7
	ACDC	60.8%	67.0%	+6.2	ACDC	ACDC	74.6%	69.9%	−4.7
	M&Ms	81.3%	82.2%	+0.9	98.84% recall	M&Ms	67.6%	47.4%	−20.2
	all	65.4%	72.6%	+7.2		all	56.6%	48.3%	−8.3
LVSC 99.75% recall	LVSC	74.3%	70.8%	−3.5		LVSC	76.5%	77.5%	+1.0
	ACDC	75.7%	79.7%	+4.0	all	ACDC	85.9%	86.8%	+0.9
	M&Ms	67.0%	69.5%	+2.5	99.69% recall	M&Ms	77.2%	83.0%	+5.8
	all	74.5%	72.7%	−1.8		all	77.7%	79.4%	+1.7

Table 3. Contingency matrix considering all voxels in subsets *all-train* and *all-test*, refer to Table 2. A corrected McNemar's chi-squared test comparing CNN without *FMF* (Base) versus with *FMF* asserted significantly different predictions, with $p < .001$.

	FMF (T)	FMF (F)	
Base (T)	6,041,403	2,543,222	$\chi^2 = 885305.1$
Base (F)	818,157	709,767,778	$p < 10^{-302}$

Table 4. Mean training time for CNN without *FMF* (Base) and with *FMF*, concerning 30 epochs. In all the cases, *FMF* accelerated the training speed, as the CNN processes less data.

Dataset	ACDC	M&Ms	LVSC	all
Base	21.0s	93.2s	254.7s	427.2s
FMF	11.6s	37.6s	95.1s	170.3s
Speed-up	1.81x	2.48x	2.68x	2.51x

With dataset LVSC used for training, we observed a lower performance when validating over itself (-3.5), but an improvement when validating with ACDC and M&Ms (+4.0 and +2.5, respectively), which corresponds to the overall mean performance of +0.3. With ACDC used for training, we did not observe performance improvements, with a mean decrease of 8.3% when considering all the datasets—it is noticeable that the Recall of RoIs had the smallest performance for this particular dataset. Finally, when using all the datasets for training, we observed a mean performance increase of +2.35 concerning all the datasets.

We applied the corrected McNemar's chi-squared test from package statsmodels 0.12.2, whose results asserted that the Base predictions and predictions after *FMF* were significantly different, with $p < .001$ (see Table 3). From a practical perspective, we also compared the training speed, which is paramount to larger experiments such as hyper-parameter search. Table 4 shows the training time of the networks without and with method *FMF*; in all the cases our method improved the training speed by 150%.
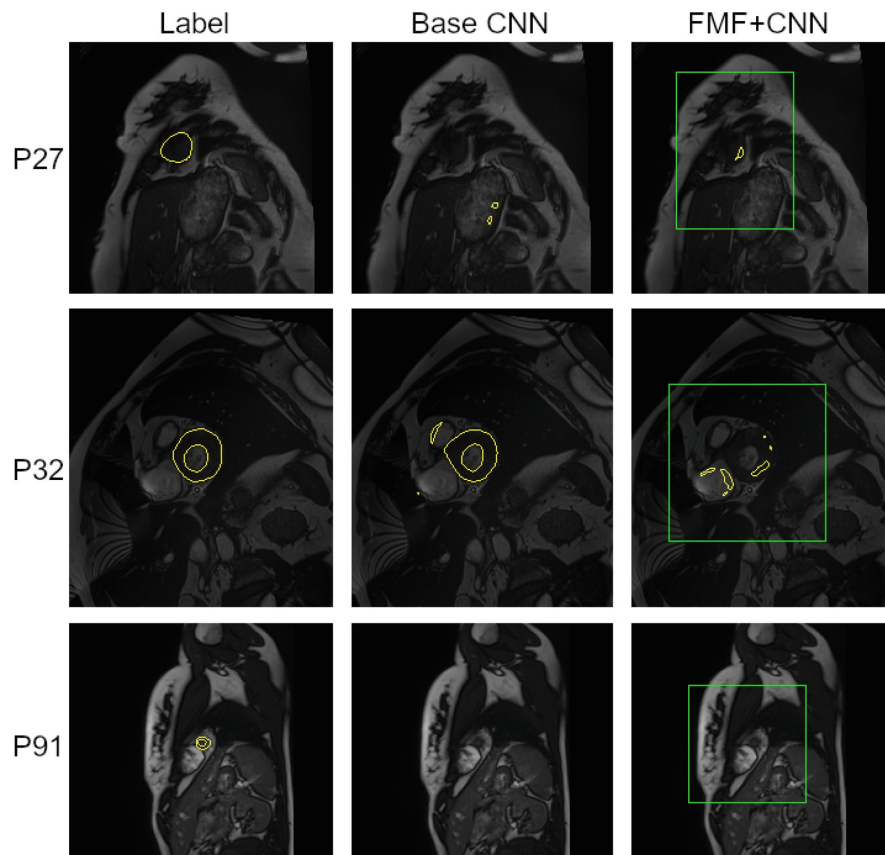
Fig. 4. Slices with DSC<0.2 from patients P27, 32 and 91. Columns are labels, predictions by base CNN, and predictions by *FMF* (rectangles) plus CNN (contours).

5   Discussion

Based on our results, the main achievement of method *FMF* is the ability to identify and crop the RoIs with a very small error (Recall = 99.69%). Furthermore,

by cropping the RoIs before the CNN training, we observed that the training process was remarkably accelerated, increasing the training speed by 150%. When considering metric DSC, we observed significant improvements when considering the entire bundle of experiments, as presented in Table 2.

For the majority of the datasets, all but ACDC used for training, we observed significant improvements. This inefficacy with ACDC is possibly related to imaging artifacts and standard operating procedures used to label the ACDC images (refer to Fig. 4). In general, *FMF* is able to automatically focus on the correct region and guide the segmentation CNN (Fig. 4, P27). However, in some cases, the CNN lost track of the LV when fed with a cropped RoI (P32). In a few cases, the CNN is unable to detect anything, even when fed with a centered RoI (P91). This is a known problem when segmenting the cardiac apex. These findings should be investigated in future works.

Overall, method *FMF* demonstrated significant improvements in the task of RoI detection; from the results, it became evident that the performance depends on the training dataset and on the network model. The main characteristic is image quality; format and intensities should be similar across datasets, and labeling standard operating procedures should be compatible. An extended preprocessing has the potential to overcome such issues across heterogeneous datasets; possibly, a dynamically adjusted module would be insensitive to data noise/variation.

6   Conclusions

In this paper, we proposed method *FMF*, a novel approach based on convolution operations and on the use of a radial basis function to detect the RoI in cardiac magnetic resonance images. We validated *FMF* with a U-Net CNN comparing our results to those of the canonical U-Net and of the *FMF*-CNN in three public reference datasets. According to our results, *FMF* was able to crop 99.69% (Recall metric) of the RoI voxels in all the datasets, being suitable to preprocess the data for CNN segmentation. *FMF* accelerated the training process by 150%, and also increased Sørensen-Dice coefficient in the majority of our test cases ($p < .001$). Further improvement is possible by extended preprocessing of the training datasets, and by the use of more advanced CNNs that support a fine-tuning to deal with the specificities of more challenging datasets.

As future work, we intend to expand the *FMF* methodology considering more possibilities: (1) use other networks in the segmentation step, such as 3D U-Net or Feature Pyramid Networks, validating versus other state-of-the-art networks; (2) embed this method as the first layer of a CNN, so that the parameters are adjusted automatically; (3) experiment with more datasets, preprocessing, and augmentation methods; (4) test different feature extractors for the static visual features, such as gray-level co-occurrence matrices and Gabor filters; other motion estimation methods, and radial basis functions.

References

1. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502

2. Earls, J.P., Ho, V.B., Foo, T.K., Castillo, E., Flamm, S.D.: Cardiac MRI: recent progress and continued challenges. J. Magn. Reson. Imaging **16**(2), 111–127 (2002). https://doi.org/10.1002/jmri.10154

3. Fallat, S.M., Johnson, C.R.: Hadamard powers and totally positive matrices. Linear Algebra Appl. **423**(2–3), 420–427 (2007). https://doi.org/10.1016/j.laa.2007.01.012

4. Guo, F., Ng, M., Wright, G.: Cardiac MRI left ventricle segmentation and quantification: a framework combining U-Net and continuous max-flow. In: Pop, M., Sermesant, M., Zhao, J., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 450–458. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_48

5. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015). https://doi.org/10.1109/TPAMI.2015.238982

6. Horn, B.K., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)

7. Khened, M., Alex, V., Krishnamurthi, G.: Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 140–151. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_15

8. Majdisova, Z., Skala, V.: Radial basis function approximations: comparison and applications. Appl. Math. Model. **51**, 728–743 (2017). https://doi.org/10.1016/j.apm.2017.07.033

9. Peng, P., Lekadir, K., Gooya, A., Shao, L., Petersen, S.E., Frangi, A.F.: A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. Magn. Reson. Mater. Phys., Biol. Med. **29**(2), 155–195 (2016). https://doi.org/10.1007/s10334-015-0521-4

10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

11. Saber, M., Abdelrauof, D., Elattar, M.: Multi-center, multi-vendor, and multi-disease cardiac image segmentation using scale-independent multi-gate UNET. In: Puyol Anton, E., et al. (eds.) STACOM 2020. LNCS, vol. 12592, pp. 259–268. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68107-4_26

12. Seraphim, A., Knott, K.D., Augusto, J., Bhuva, A.N., Manisty, C., Moon, J.C.: Quantitative cardiac MRI. J. Magn. Reson. Imaging **51**(3), 693–711 (2020). https://doi.org/10.1002/jmri.26789

13. Suinesiaputra, A., et al.: A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. Med. Image Anal. **18**(1), 50–62 (2014). https://doi.org/10.1016/j.media.2013.09.001

14. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR 2015, pp. 1–9. IEEE (2015). https://doi.org/10.1109/CVPR.2015.7298594
15. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) NIPS 2017. Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates Inc, Red Hook (2017)
16. Wu, B., Fang, Y., Lai, X.: Left ventricle automatic segmentation in cardiac MRI using a combined CNN and U-net approach. Comput. Med. Imaging Graph. **82**, 101719 (2020). https://doi.org/10.1016/j.compmedimag.2020.101719
17. Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S.: Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. **43**, 54–65 (2018). https://doi.org/10.1016/j.media.2017.09.005
18. Xue, W., Lum, A., Mercado, A., Landis, M., Warrington, J., Li, S.: Full quantification of left ventricle via deep multitask learning network respecting intra- and inter-task relatedness. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 276–284. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_32
19. Yan, W., Wang, Y., Li, Z., van der Geest, R.J., Tao, Q.: Left ventricle segmentation via optical-flow-net from short-axis cine MRI: preserving the temporal coherence of cardiac motion. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 613–621. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_70

5

CONCLUSION

This thesis contributes to a thorough process for analyzing real data in the medical domain from the perspective of computer science applied to (A) Clinical Data Warehouses, (B) Skin Lesion Identification, and (C) Cardiac MRI Analysis. In this project we prepared a data warehouse from real clinical data, then explored deep learning analysis in two medical specialties: dermatology and cardiology, considering both high-level diagnostic concepts and low-level morphology concepts. The topics are essential to assert the following:

> **General Hypothesis**.
> *Deep Learning is able to analyze real medical data, adding to the clinical information about a patient's condition and possible outcomes. That ability works both for learning high-level concepts such as diagnostics from curated clinical data warehouses and low-level concepts such as motion patterns in medical images.*

The data warehouse application comprises a specialized database for structured medical data. It stores coalesced patient records with clinical meaning, that is, using standard terms acknowledged by medical literature. This information is often curated in statistical models of human anatomy and laboratory specimens. The data warehouse thus serves as a support tool for understanding persons' health, the possible outcomes and treatments.

The skin lesion and cardiac MRI applications access unstructured data in the form of images from skin and heart, which are analyzed to detect patterns related to dimensions, shape, texture and color. Those patterns are important characteristics for determination of diseases and structured information about the organs and tissues in evaluation. They serve as basis for clinical measurements such as lesion diameter or heart cavity volumes. Measurements such as diameter and volumes are input variables used for clinical risk models, for example, in cancer screening and heart failure diagnosis.

5.1 Further discussion

5.1.1 *Clinical Data Warehouses*

Chapter 2 described the ETL process that built a data warehouse based on InCor's clinical database in accordance to the OMOP CDM standard. InCor CDM was able to reproduce cohort selections from previous studies, and also included additional medical data that opened doors for future projects. The main point observed when defining cohorts using Atlas instead of Pauá was that their queries are slightly off due to undefined constraints, such as window gap size and number rounding. This required manual adjustment of some parameters to find the most agreeable result. Another point is that the source database is ever changing, that is, some patients had record updates or deletions, which makes it impossible to reproduce cohorts exactly even if we created Pauá views again. However, this characteristic is desirable in the clinical research context, if the database was updated to fix wrong data inputs or add details. It was also discovered that the OMOP CDM did not include raw medical data in their model, requiring links to an external system such as a PACS using a separate key. In our case, we used an additional table to link the CDM occurrence ids to the PACS ids.

5.1.2 *Skin Lesion Identification*

Chapter 3 demonstrated an extension of the DermaDL melanoma detection CNN to a more general task of classifying skin lesions. A new code base was implemented in Tensorflow to help understanding the CNN block formulation, which also validated that the model does not depend on hidden optimizations from a specific DL library. We observed that the new CNN was more difficult to train as the initial results had poor metrics, and therefore additional work would be required. The first concern was with the strong dataset imbalance both between classes and between datasets, and also that there were many classes with very few examples. When futher splitting to multiple classes, the imbalance increased even more. A second concern that aroused was if different preprocessing configurations and approaches could affect class separation. We closed those gaps by regrouping the dataset classes in five classes, we provided colorized versions of image equalization, and finally augmented the dataset a second time to alleviate the class imbalance. In this contribution we observed that even if the CNN has a complex structure, it has much less weights compared to other DL models, making the architecture suitable for energy-constrained devices such as mobile phones or portable devices.

5.1.3 *Cardiac MRI Analysis*

Chapter 4 contributed with a method to focus on the heart movements and use that information to crop the heart region before applying a CNN. The method was carefully defined with convolution operations so future implementations could be added to a CNN architecture if necessary, and possibly allowing for automatic weight adjustments during the network training

procedure. The experiments showed promising, but they also exposed some difficult cases from the ACDC dataset. The FMF method correctly identified the heart location in those cases, but the CNN segmentation was worse than before. This suggests that some types of CMR images may have such unexpected characteristics so that the CNNs tries to learn secondary features from the regions outside the RoI. It should be noted that the U-Net used after FMF was not fine-tuned, and no augmentation was done to the dataset, as the study focused on FMF.

5.2   Open problems

InCor CDM is just one of multiple healthcare centers that has adopted an standardized data warehouse. The main benefit of using an standard model is sharing of materials and methods, which facilitates interoperability between healthcare systems and reproduction of clinical research protocols. If the data follows the same model, the software and methods can be reused with minimal changes. However, many systems are designed with a multitude of non-relational database subsystems, specially for cost or time constraints.

This may pose a challenge when the institution attempts to organize a data warehouse following a relational model. Even further, some institutions may use internal medical vocabularies that are not compatible with public international standards such as the ICD-10 diagnostic code system. Tito *et al.* (2020) proposed a fully automatic approach for data analysis in unstructured datasets, while Lima *et al.* (2021) proposed to use intermediate data models that could be combined into a structured data lake. However, the best choice still depends on the institution research goals, which can be both geared towards understanding unknown data or strictly following clinical research protocol with a clearly defined data model.

Regarding the DL applications in the medical domain, current state-of-the-art already surpassed human performance in experiments, as reported by Esteva *et al.* (2017). However, there is a long road of building confidence in the computerized decisions that involve healthcare, such as explaining all steps of the algorithms to the medical specialists, acquiring approval from government agencies and obtaining the trust of the general public. It is very challenging to explain DL concepts and reasoning in a mathematically precise manner for the specialists and that is also understandable and intuitive to everyone else. For that reason the medical literature still makes heavy usage of feature engineering and linear models, due to the inherent explainability and mathematical simplicity.

The DL community has made several attempts of explaining CNN behaviour by visualizing error gradients, such as in the work of Selvaraju *et al.* (2020) and similar approaches. This visual explainability framework also helps researchers to understand how CNN architectures guide the objective function with attention mechanisms that are used in other DL models such as in the work of Vaswani *et al.* (2017). Another common reported visualization is the learned kernels, but all of these approaches are still too technical.

Another concern regarding DL research intersects the data interoperability scenario, because there is a lack of data variability in medical imaging datasets that is currently being addressed by the scientific community. For example, considering the CMR imaging datasets used in Chapter 4, we observed that literature's datasets prior to the M&Ms challenge of Campello *et al.* (2021) were efforts done by leading healthcare institutions that often had small variability of MRI machines and protocols, and were focused on particular diseases or learning tasks. It is completely perfect for motivating scientific research and development of DL models, but caution must be taken when applying the resulting models to real data acquired in other hospitals, with different MRI equipment, protocols and population characteristics.

5.3   Scientific production

This section provides a list of Lima's publications from 2018 to 2023, which were divided into first-authored publications (main research line) and co-authored publications (collaborations with research groups). More information about these publications and their full texts is available in his Lattes, ResearchGate, and ORCID profiles.

5.3.1   *First-authored publications*

— Journal Articles —

1 - Lima *et al.* (2021). *DermaDL: Advanced convolutional neural networks for computer-aided skin-lesion classification.* SN Computer Science 2(253), Springer Nature.

2 - Lima *et al.* (2020). *A COVID-19 surveillance platform to monitor risk of infection based on a machine learning model.* Journal of Health Informatics 12(s1), SBIS.

3 - Lima *et al.* (2019). *Transforming two decades of ePR data to OMOP CDM for clinical research.* Studies in Health Technology and Informatics 264, IOS Press.

— Book Chapters —

4 - Lima *et al.* (2021). *Evaluating the risk of COVID-19 infection based on machine learning of symptoms and conditions versus laboratory methods.* Coleção desafios das engenharias - Engenharia de computação 3, Atena.

— Conference Proceedings —

5 - Lima *et al.* (2022). *Full Motion Focus: Convolutional Module for Improved Left Ventricle Segmentation over 4D MRI.* In: Proceedings of Image Analysis and Processing – ICIAP 2022.

6 - Lima *et al.* (2021). *Uma proposta de data lake para pesquisa em saúde a partir de data pools multicêntricos interoperáveis*. In: Anais do XXXVI SBBD, SBC.

5.3.2   Co-authored publications

— Journal Articles —

7 - Carvalho *et al.* (2023). *Long-term respiratory follow-up of ICU hospitalized COVID-19 patients: prospective cohort study*. PLOS ONE 18(1), Public Library of Science.

8 - Carvalho *et al.* (2022). *Chronic lung lesions in COVID-19 survivors: predictive clinical model*. BMJ Open 12(6), British Medical Journal Publishing Group.

9 - Linhares *et al.* (2022) *ClinicalPath: a Visualization tool to Improve the Evaluation of Electronic Health Records in Clinical Decision-Making*. IEEE Transactions on Visualization and Computer Graphics, IEEE.

10 - Toledo *et al.* (2021). *Study of CNN Capacity Applied to Left Ventricle Segmentation in Cardiac MRI*. SN Computer Science 2(480), Springer Nature.

11 - Costa *et al.* (2022). *Blood pressure estimation from photoplethysmography by considering intra- and inter-subject variabilities: guidelines for a fair assessment*. Submitted to IEEE Access.

— Conference Proceedings —

12 - Linhares *et al.* (2021). *I-CovidVis – A Visual Analytics Tool for Interoperable Healthcare Databases using Graphs*. In: Proceedings of IEEE 34th CBMS.

13 - Clementino *et al.* (2020). *Bag-of-Attributes Representation: A Vector Space Model for Electronic Health Records Analysis in OMOP*. In: Proceedings of IEEE 33rd CBMS.

14 - Pereira *et al.* (2020). *Fully Automated Quantification of Cardiac Indices from Cine MRI Using a Combination of Convolution Neural Networks*. In: Proceedings of IEEE 42nd EMBC.

15 - Vasconcelos *et al.* (2018). *TendeR-Sims - Similarity Retrieval System for Public Tenders*. In: Proceedings of 20th ICEIS, SciTePress.

BIBLIOGRAPHY

ABRAHÃO, M. T.; NOBRE, M. R.; GUTIERREZ, M. A. Seleção sistemática de coortes em bases de dados assistenciais. **Value in Health**, v. 20, n. 9, p. A860–A861, 2017. Available: <https://doi.org/10.1016/j.jval.2017.08.2484>. Citation on page 19.

AGRAWAL, R.; AILAMAKI, A.; BERNSTEIN, P. A.; BREWER, E. A.; CAREY, M. J.; CHAUD-HURI, S.; DOAN, A.; FLORESCU, D.; FRANKLIN, M. J.; GARCIA-MOLINA, H.; GEHRKE, J.; GRUENWALD, L.; HAAS, L. M.; HALEVY, A. Y.; HELLERSTEIN, J. M.; IOANNIDIS, Y. E.; KORTH, H. F.; KOSSMANN, D.; MADDEN, S.; MAGOULAS, R.; OOI, B. C.; O'REILLY, T.; RAMAKRISHNAN, R.; SARAWAGI, S.; STONEBRAKER, M.; SZALAY, A. S.; WEIKUM, G. The Claremont report on database research. **SIGMOD Record**, Association for Computing Machinery, New York, NY, USA, v. 37, n. 3, p. 9–19, sep 2008. ISSN 0163-5808. Available: <https://doi.org/10.1145/1462571.1462573>. Citations on pages 17 and 18.

American Cancer Society. **Cancer Facts and Figures**. 2022. Accessed: 20-Oct-2022. Available: <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html>. Citation on page 35.

CAMPELLO, V. M.; GKONTRA, P.; IZQUIERDO, C.; MARTÍN-ISLA, C.; SOJOUDI, A.; FULL, P. M.; MAIER-HEIN, K.; ZHANG, Y.; HE, Z.; MA, J.; PARREÑO, M.; ALBIOL, A.; KONG, F.; SHADDEN, S. C.; ACERO, J. C.; SUNDARESAN, V.; SABER, M.; ELATTAR, M.; LI, H.; MENZE, B.; KHADER, F.; HAARBURGER, C.; SCANNELL, C. M.; VETA, M.; CARSCADDEN, A.; PUNITHAKUMAR, K.; LIU, X.; TSAFTARIS, S. A.; HUANG, X.; YANG, X.; LI, L.; ZHUANG, X.; VILADÉS, D.; DESCALZO, M. L.; GUALA, A.; MURA, L. L.; FRIEDRICH, M. G.; GARG, R.; LEBEL, J.; HENRIQUES, F.; KARAKAS, M.; ÇAVUŞ, E.; PETERSEN, S. E.; ESCALERA, S.; SEGUÍ, S.; RODRÍGUEZ-PALOMARES, J. F.; LEKADIR, K. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. **IEEE Transactions on Medical Imaging**, v. 40, n. 12, p. 3543–3554, 2021. Available: <https://doi.org/10.1109/TMI.2021.3090082>. Citation on page 70.

CARVALHO, C.; CHATE, R.; SAWAMURA, M.; GARCIA, M.; LAMAS, C.; CARDENAS, D.; LIMA, D.; SCUDELLER, P.; SALGE, J.; NOMURA, C.; GUTIERREZ, M. Chronic lung lesions in COVID-19 survivors: predictive clinical model. **BMJ Open**, British Medical Journal Publishing Group, v. 12, n. 6, 2022. ISSN 2044-6055. Available: <http://doi.org/10.1136/bmjopen-2021-059110>. Citation on page 71.

CARVALHO, C. R. R.; LAMAS, C. A.; CHATE, R. C.; SALGE, J. M.; SAWAMURA, M. V. Y.; ALBUQUERQUE, A. L. P.; JUNIOR, C. T.; LIMA, D. M.; GARCIA, M. L.; SCUDELLER, P. G.; NOMURA, C. H.; GUTIERREZ, M. A.; BALDI, B. G. Long-term respiratory follow-up of ICU hospitalized COVID-19 patients: prospective cohort study. **PLOS ONE**, v. 18, n. 1, p. 1–16, 2023. Available: <https://doi.org/10.1371/journal.pone.0280567>. Citation on page 71.

CLEMENTINO, J. M.; BONES, C. C.; FAIÇAL, B. S.; LINARES, O. C.; LIMA, D. M.; GUTIERREZ, M. A.; TRAINA, C.; TRAINA, A. J. M. Bag-of-attributes representation: A vector space model for electronic health records analysis in OMOP. In: **33rd International Symposium on**

Computer-Based Medical Systems (CBMS). IEEE, 2020. p. 197–202. ISBN 978-1-7281-9429-5. ISSN 2372-9198. Available: <http://doi.org/10.1109/CBMS49503.2020.00045>. Citations on pages 24, 28, and 71.

COBAS, R.; RODACKI, M.; GIACAGLIA, L.; CALLIARI, L.; NORONHA, R.; VALERIO, C.; CUSTÓDIO, J.; SANTOS, R.; ZAJDENVERG, L.; GABBAY, G.; BERCOLUCI, M. **Diagnóstico do diabetes e rastreamento do diabetes tipo 2. Diretriz Oficial da Sociedade Brasileira de Diabetes**. SBD, 2022. Accessed: 18-Oct-2022. Available: <https://diretriz.diabetes.org.br/diagnostico-e-rastreamento-do-diabetes-tipo-2/>. Citation on page 27.

ESTEVA, A.; KUPREL, B.; NOVOA, R. A.; KO, J.; SWETTER, S. M.; BLAU, H. M.; THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, p. 115–118, 2017. ISSN 1476-4687. Available: <https://doi.org/10.1038/nature21056>. Citations on pages 18 and 69.

FARIAS, J. S.; GUIMARAES, T. D. A.; VARGAS, E. R. D.; ALBUQUERQUE, P. H. M. Adoção de prontuário eletrônico do paciente em hospitais universitários de Brasil e Espanha: a percepção de profissionais de saúde. **Revista de Administração Pública**, v. 45, n. 5, p. 1303–1326, 2011. Available: <https://doi.org/10.1590/S0034-76122011000500004>. Citation on page 19.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. In: **The Second International Conference on Knowledge Discovery and Data Mining (KDD-96)**. AAAI, 1996. p. 82–88. Accessed: 05-Dec-2022. Available: <https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf?utm_campaign=ml4devs-newsletter&utm_medium=email&utm_source=Revue%20newsletter>. Citation on page 22.

FOLEY, T. A.; MANKAD, S. V.; ANAVEKAR, N. S.; BONNICHSEN, C. R.; MORRIS, M. F.; MILLER, T. D.; ARAOZ, P. A. Measuring left ventricular ejection fraction - techniques and potential pitfalls. **European Cardiology**, v. 8, n. 2, p. 108–14, 2012. Available: <https://doi.org/10.15420/ecr.2012.8.2.108>. Citation on page 52.

FURUIE, S. S.; REBELO, M. S.; MORENO, R. A.; SANTOS, M.; BERTOZZO, N.; MOTTA, G. H. M. B.; PIRES, F. A.; GUTIERREZ, M. A. Managing medical images and clinical information: InCor's experience. **IEEE Transactions on Information Technology in Biomedicine**, v. 11, n. 1, p. 17–24, 2007. Available: <https://doi.org/10.1109/TITB.2006.879588>. Citation on page 18.

GARDNER, R. M.; PRYOR, T. A.; WARNER, H. R. The HELP hospital information system: update 1998. **International journal of medical informatics**, Elsevier, v. 54, n. 3, p. 169–182, 1999. Available: <https://doi.org/10.1016/S1386-5056(99)00013-1>. Citation on page 18.

GARETS, D.; DAVIS, M. Electronic patient records. **Healthcare Informatics Online**, v. 4, 2005. Accessed: 20-Oct-2022. Available: <http://providersedge.com/ehdocs/ehr_articles/Electronic_Patient_Records-EMRs_and_EHRs.pdf>. Citation on page 18.

GEYER, F.; CARLE, G. Learning and generating distributed routing protocols using graph-based deep learning. In: **Proceedings of the 2018 Workshop on Big Data Analytics and Machine Learning for Data Communication Networks (Big-DAMA '18)**. New York, NY, USA: Association for Computing Machinery, 2018. p. 40–45. ISBN 9781450359047. Available: <https://doi.org/10.1145/3229607.3229610>. Citation on page 18.

GOLDBERG, D. G.; KUZEL, A. J.; FENG, L. B.; DESHAZO, J. P.; LOVE, L. E. EHRs in primary care practices: benefits, challenges, and successful strategies. **The American Journal of Managed Care**, v. 18, n. 2, p. e48–54, 2012. Accessed: 05-Dec-2022. Available: <https://www.ajmc.com/view/ehrs-in-primary-care-practices-benefits-challenges-and-successful-strategies>. Citation on page 19.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. MIT press, 2016. 800 p. Accessed: 05-Dec-2022. ISBN 0262035618. Available: <https://www.deeplearningbook.org/>. Citation on page 20.

GOTCH, F.; HORSLEY, V. VI Croonian lecture – on the mammalian nervous system, its functions, and their localisation by an electrical method. **Philosophical Transactions of the Royal Society of London: B.**, The Society, 1892. Citation on page 21.

HENRY, J.; PYLYPCHUK, Y.; SEARCY, T.; PATEL, V. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2015. **ONC Data Brief**, The Office of the National Coordinator for Health Information Technology, v. 35, 2016. Accessed: 20-Oct-2022. Available: <https://www.healthit.gov/data/data-briefs/adoption-electronic-health-record-systems-among-us-non-federal-acute-care-1>. Citations on pages 18 and 19.

HRIPCSAK, G.; DUKE, J. D.; SHAH, N. H.; REICH, C. G.; HUSER, V.; SCHUEMIE, M. J.; SUCHARD, M. A.; PARK, R. W.; WONG, I. C. K.; RIJNBEEK, P. R.; LEI, J. van der; PRATT, N.; NORÉN, G. N.; LI, Y.-C.; STANG, P. E.; MADIGAN, D.; RYAN, P. B. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. **MEDINFO 2015: eHealth-enabled Health**, v. 216, p. 574–578, 2015. Available: <https://doi.org/10.3233/978-1-61499-564-7-574>. Citation on page 18.

INMON, W. H. **Building the data warehouse**. Indianapolis: Wiley Publishing, 2005. ISSN 0-7645-9944-6. ISBN 978-0-7645-9944-6. Citation on page 18.

Instituto Nacional de Câncer do Brasil. **Estimativa 2020: incidência de câncer no Brasil**. 2020. Accessed: 20-Oct-2022. Available: <https://www.inca.gov.br/estimativa>. Citation on page 35.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, p. 84 – 90, 2012. Available: <https://doi.org/10.1145/3065386>. Citations on pages 17 and 21.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, p. 436–444, 2015. ISSN 1476-4687. Available: <https://doi.org/10.1038/nature14539>. Citation on page 17.

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. Citation on page 21.

LIMA, D.; MORENO, R.; PIRES, F.; GUTIERREZ, M. Uma proposta de data lake para pesquisa em saúde a partir de data pools multicêntricos interoperáveis. In: **Anais do XXXVI Simpósio Brasileiro de Bancos de Dados (SBBD)**. Porto Alegre, RS, Brasil: SBC, 2021. p. 367–372. ISSN 2763-8979. Available: <http://doi.org/10.5753/sbbd.2021.17900>. Citations on pages 24, 69, and 71.

LIMA, D. M.; GRAVES, C. V.; GUTIERREZ, M. A.; BRANDOLI, B.; RODRIGUES, J. F. Full motion focus: Convolutional module for improved left ventricle segmentation over 4D MRI. In: SCLAROFF, S.; DISTANTE, C.; LEO, M.; FARINELLA, G. M.; TOMBARI, F. (Ed.). **Image Analysis and Processing – ICIAP 2022**. Cham: Springer International Publishing, 2022. (Lecture Notes in Computer Science, v. 13231), p. 438–450. ISBN 978-3-031-06427-2. Available: <http://doi.org/10.1007/978-3-031-06427-2_37>. Citations on pages 25, 51, 52, and 70.

LIMA, D. M.; MORENO, R. A.; REBELO, M. S.; KRIEGER, J. E.; GUTIERREZ, M. A. A COVID-19 surveillance platform to monitor risk of infection based on a machine learning model. In: **XVII Congresso Brasileiro de Informática em Saúde - CBIS 2020**. SBIS, 2020. (Journal of Health Informatics, v. 12). ISSN 2175-4411. Accessed: 05-Dec-2022. Available: <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/839>. Citation on page 70.

LIMA, D. M.; RODRIGUES JR., J. F.; BRANDOLI, B.; GOEURIOT, L.; AMER-YAHIA, S. DermaDL: Advanced convolutional neural networks for computer-aided skin-lesion classification. **SN Computer Science**, Springer Nature, v. 2, n. 253, p. 1–13, 2021. ISSN 2661-8907. Available: <http://doi.org/10.1007/s42979-021-00641-5>. Citations on pages 25, 35, 36, and 70.

LIMA, D. M.; RODRIGUES JR., J. F.; TRAINA, A. J. M.; PIRES, F. A.; GUTIERREZ, M. A. Transforming two decades of ePR data to OMOP CDM for clinical research. In: **MEDINFO 2019: Health and Wellbeing e-Networks for All**. Lyon, France: IOS Press, 2019. (Studies in Health Technology and Informatics, v. 264), p. 233–237. ISSN 1879-8365. Available: <http://doi.org/10.3233/SHTI190218>. Citations on pages 25, 27, 28, and 70.

LIMA, D. M.; SÁ, J. H. G.; MORENO, R. A.; REBELO, M. F. S.; KRIEGER, J. E.; GUTIERREZ, M. A. Evaluating the risk of COVID-19 infection based on machine learning of symptoms and conditions versus laboratory methods. In: ____. **Engenharia de Computação 3**. Ponta Grossa, PR: Atena, 2021. (Coleção desafios das engenharias), p. 1–15. ISBN 978-65-5983-619-2. Available: <http://doi.org/10.22533/at.ed.1922129111>. Citations on pages 19 and 70.

LINHARES, C. D. G.; LIMA, D. M.; BONES, C. C.; REBELO, M. F. S.; GUTIERREZ, M. A.; TRAINA, C.; TRAINA, A. J. M. I-CovidVis – a visual analytics tool for interoperable healthcare databases using graphs. In: **34th International Symposium on Computer-Based Medical Systems (CBMS)**. IEEE, 2021. p. 125–130. ISBN 978-1-6654-4121-6. ISSN 2372-9198. Available: <http://doi.org/10.1109/CBMS52027.2021.00059>. Citation on page 71.

LINHARES, C. D. G.; LIMA, D. M.; PONCIANO, J. R.; OLIVATTO, M. M.; GUTIERREZ, M. A.; POCO, J.; TRAINA, C.; TRAINA, A. J. M. ClinicalPath: a visualization tool to improve the evaluation of electronic health records in clinical decision-making. **IEEE Transactions on Visualization and Computer Graphics**, p. 1–1, 2022. Available: <http://doi.org/10.1109/TVCG.2022.3175626>. Citations on pages 24 and 71.

MACEDO, B. R. de; GARCIA, M. V. F.; GARCIA, M. L.; VOLPE, M.; SOUSA, M. L. de A.; AMARAL, T. F.; GUTIERREZ, M. A.; BARBOSA, A. P.; SCUDELLER, P. G.; CARUSO, P.; CARVALHO, C. R. R. Implantação de telemedicina de terapia intensiva durante a pandemia de COVID-19. **Jornal Brasileiro de Pneumologia**, v. 47, n. 2, 2021. Available: <https://doi.org/10.36416/1806-3756/e20200545>. Citation on page 19.

MCCULLOCH, W. S.; PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, v. 5, p. 115–133, 1943. Citation on page 21.

MONTERA, M. W.; MARCONDES-BRAGA, F. G.; SIMÕES, M. V.; MOURA, L. A. Z.; FERNAN-DES, F.; MANGINE, S.; OLIVEIRA JÚNIOR, A. C. d.; SOUZA, A. L. A. d. A. G. d.; IANNI, B. M.; ROCHITTE, C. E.; MESQUITA, C. T.; AZEVEDO FILHO, C. F. d.; FREITAS, D. C. d. A.; MELO, D. T. P. d.; BOCCHI, E. A.; HOROWITZ, E. S. K.; MESQUITA, E. T.; OLIVEIRA, G. H.; VILLA-CORTA, H.; ROSSI NETO, J. M.; BARBOSA, J. M. B.; FIGUEIREDO NETO, J. A. d.; LUIZ, L. F.; HAJJAR, L. A.; BECK-DA-SILVA, L.; CAMPOS, L. A. d. A.; DANZMANN, L. C.; BITTENCOURT, M. I.; GARCIA, M. I.; AVILA, M. S.; CLAUSELL, N. O.; OLIVEIRA JR, N. A. d.; SILVESTRE, O. M.; SOUZA, O. F. d.; MOURILHE-ROCHA, R.; KALIL FILHO, R.; AL-KINDI, S. G.; RASSI, S.; ALVES, S. M. M.; FERREIRA, S. M. A.; RIZK, S. I.; MATTOS, T. A. C.; BARZILAI, V.; MARTINS, W. d. A.; SCHULTHEISS, H.-P. Brazilian society of cardiology guideline on myocarditis – 2022. **Arquivos Brasileiros de Cardiologia**, Sociedade Brasileira de Cardiologia, v. 119, n. 1, p. 143–211, 2022. Available: <http://doi.org/10.36660/abc.20220412>. Citation on page 52.

OWEIS, N.; OWAIS, S.; GEORGE, W.; SULIMAN, M.; SNASEL, V. A survey on big data, mining: (tools, techniques, applications and notable uses). In: **Intelligent Data Analysis and Applications**. Springer, 2015. (Advances in Intelligent Systems and Computing, v. 370). Available: <https://doi.org/10.1007/978-3-319-21206-7_10>. Citation on page 17.

PENG, P.; LEKADIR, K.; GOOYA, A.; SHAO, L.; PETERSEN, S. E.; FRANGI, A. F. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. **Magnetic Resonance Materials in Physics, Biology and Medicine**, v. 19, p. 155–195, 2016. Available: <http://doi.org/10.1007/s10334-015-0521-4>. Citation on page 51.

PEREIRA, R. F.; REBELO, M. S.; MORENO, R. A.; MARCO, A. G.; LIMA, D. M.; ARRUDA, M. A. F.; KRIEGER, J. E.; GUTIERREZ, M. A. Fully automated quantification of cardiac indices from cine MRI using a combination of convolution neural networks. In: **42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)**. IEEE, 2020. p. 1221–1224. ISBN 978-1-7281-1990-8. ISSN 2694-0604. Available: <http://doi.org/10.1109/EMBC44109.2020.9176166>. Citation on page 71.

RAISSI, M. Deep hidden physics models: Deep learning of nonlinear partial differential equations. **Journal of Machine Learning Research**, v. 19, p. 1–24, 2018. Available: <https://doi.org/10.5555/3291125.3291150>. Citation on page 18.

RASHEVSKY, N. Mathematical biophysics. **Nature**, 1935. Citation on page 21.

RODRIGUES, J.; BRANDOLI, B.; AMER-YAHIA, S. DermaDL: Advanced convolutional neural networks for automated melanoma detection. In: **33rd International Symposium on Computer-Based Medical Systems (CBMS)**. IEEE Press, 2020. p. 504–509. ISBN 978-1-7281-9429-5. ISSN 2372-9198. Available: <https://ieeexplore.ieee.org/document/9182826>. Citation on page 35.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N.; HORNEGGER, J.; WELLS, W.; FRANGI, A. (Ed.). **International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)**. Cham: Springer, 2015. (Lecture Notes in Computer Science, v. 9351), p. 234–241. Available: <https://doi.org/10.1007/978-3-319-24574-4_28>. Citation on page 23.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958. Available: <https://doi.org/10.1037/h0042519>. Citation on page 21.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. **International Journal of Computer Vision**, Springer, v. 128, p. 336–359, 2020. Available: <https://doi.org/10.1007/s11263-019-01228-7>. Citation on page 69.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VAN-HOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE Press, 2015. p. 1–9. ISSN 1063-6919. Available: <http://doi.org/10.1109/CVPR.2015.7298594>. Citation on page 52.

TANG, B.; PAN, Z.; YIN, K.; KHATEEB, A. Recent advances of deep learning in bioinformatics and computational biology. **Frontiers in Genetics**, v. 10, 2019. ISSN 1664-8021. Available: <https://doi.org/10.3389/fgene.2019.00214>. Citation on page 17.

TITO, L.; MOTINHA, C.; SANTIAGO, F.; OCAñA, K.; BEDO, M.; OLIVEIRA, D. de. Xi-dl: um sistema de gerência de data lake para monitoramento de dados da saúde. In: **Anais do XXXV Simpósio Brasileiro de Bancos de Dados (SBBD)**. SBC, 2020. p. 151–156. Available: <https://doi.org/10.5753/sbbd.2020.13633>. Citation on page 69.

TOLEDO, M. A. F.; LIMA, D. M.; KRIEGER, J. E.; GUTIERREZ, M. A. Study of CNN capacity applied to left ventricle segmentation in cardiac MRI. **SN Computer Science**, Springer Nature, v. 2, n. 480, 2021. ISSN 2661-8907. Available: <http://doi.org/10.1007/s42979-021-00897-x>. Citation on page 71.

TURING, A. M. On computable numbers, with an application to the entscheidungsproblem. **Journal of Mathematics**, v. 58, p. 345–363, 1936. Citation on page 21.

VASCONCELOS, G. Q.; ZABOT, G. F.; LIMA, D. M.; RODRIGUES JR., J. F.; TRAINA JR., C.; KASTER, D. S.; CORDEIRO, R. L. F. TendeR-Sims - similarity retrieval system for public tenders. In: **Proceedings of the 20th International Conference on Enterprise Information Systems - Volume 1: ICEIS**. SciTePress, 2018. p. 143–150. ISBN 978-989-758-298-1. ISSN 2184-4992. Available: <http://doi.org/10.5220/0006697601430150>. Citation on page 71.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: **31st Conference on Neural Information Processing Systems (NIPS)**. Long Beach, CA, USA: NeurIPS Press, 2017. (Advances in Neural Information Processing Systems, v. 30). Accessed: 05-Dec-2022. Available: <https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need>. Citation on page 69.

WANG, F.; CASALINO, L. P.; KHULLAR, D. Deep learning in medicine — promise, progress, and challenges. **JAMA Internal Medicine**, v. 179, n. 3, p. 293–294, 03 2019. ISSN 2168-6106. Available: <https://doi.org/10.1001/jamainternmed.2018.7117>. Citation on page 18.

World Health Organization. **International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)**. 2016. Accessed: 18-Oct-2022. Available: <https://icd.who.int/browse10/2016/en>. Citations on pages 22, 28, and 52.

WU, E.; LIU, Y. Emerging technology about GPGPU. In: **2008 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)**. IEEE, 2008. p. 618–622. Available: <https://doi.org/10.1109/APCCAS.2008.4746099>. Citation on page 17.

GLOSSARY

Anamnesis: during a medical consultation, it is the process of paying attention to the patient complaints, and inspecting the body appearance, posture, signals and symptoms, so the physician can start diagnosing what conditions or diseases the patient has.

Biostatistics: area of knowledge that applies Statistics to questions in Biology and Health Sciences.

Clinical: related or concerned with observation of a patient such as to establish a diagnostic and treatment plan.

Cohort: selection of patients that satisfy the criteria determined by the clinical researcher.

Condition: in the medical context, it is a characteristic or status of a person that is not a disease but affects the health, e.g. smoking.

Convolution: mathematical operation that generalizes multiplication to compose two functions, and is closely related to the cross-correlation.

Dermatoscopy: acquisition of skin images using a special optical device called dermatoscope.

Diagnostic: identification of a disease, problem or characteristic of a person.

Diastole: instant of the heartbeat when the left ventricle is most relaxed.

Melanoma: type of cancer that develops in the pigmented cells of the skin called melanocytes.

Segmentation: image processing task that selects regions by marking pixels one-by-one or drawing a shape around the contour of the region.

Stroke: heart movement that contracts the left ventricle and sends blood to the body.

Systole: instant of the heartbeat when the left ventricle is most contracted.

Telemedicine: healthcare practice that is executed remotely with the use of information and communication technologies.

Thresholding: process of limiting a function output to a maximum or minimum value, called the threshold.

Triage: process of identifying if the patient status is healthy, sick, urgent or critical then directing to the appropriate location.

Ventricles: the two bottom chambers of the heart, one right and one left.