CHAPTER 12

Chapter 12: Applications of AI and HPC in Health Domain

Add ORCiD here (won't appear in final printed book)

Dana Oniga https://orcid.org/0000-0003-3376-099x
Barbara Cantalupo https://orcid.org/0000-0001-7575-3902
Enzo Tartaglione https://orcid.org/0000-0003-4274-8298
Daniele Perlo https://orcid.org/0000-0001-6879-8475
Marco Grangetto https://orcid.org/0000-0002-2709-7864
Marco Aldinucci https://orcid.org /0000-0001-8788-0829
Federico Bolelli https://orcid.org/0000-0002-5299-6351
Federico Pollastri https://orcid.org/0000-0001-8036-1559
Michele Cancilla https://orcid.org/0000-0001-6106-3249
Laura Canalini https://orcid.org/0000-0001-9572-5044
Costantino Grana https://orcid.org/0000-0002-4792-2358
https://orcid.org/0000-0002-4174-3762Cristina Muñoz Alcalde https://orcid.org/0000-0003-3417-0449
Franco Alberto Cardillo https://orcid.org/0000-0003-0940-4768
Monica Florea https://orcid.org/0000-0003-0620-7558

Chapter Abstract

150 words—will not be printed in the final book, but used in the metadata for online discoverability.

12.1 Introduction

This chapter aims to show how the convergence of High Performance Computing (HPC), BigData and Artificial Intelligence (AI) (and Deep Learning in particular) can support improvements in the Health domain, by providing an overview of few use cases from the DeepHealth project. The next sections are organised as follows: Section 12.2 presents the latest technical progresses in the health field. Section 12.3 briefly describes the DeepHealth concept and introduces the DeepHealth toolkit. Section 12.4 presents general info (typical workflow, KPIs) about the DeepHealth use cases and Section 12.5 provides more details of five DeepHealth use cases. Section 12.6 highlights the value proposition offered by DeepHealth Toolkit to HealthCare providers. Section 12.7 concludes emphasizing the impact of the DeepHealth project.

12.2 Artificial Intelligence and HPC in Health Domain in 2020

The European Commission (EC) focuses on innovation in health and on technical areas addressing current health issues in relation to people's well-being, as well as on increasing the sustainability of health systems, as presented in the EU eHealth Action Plan (2014-2020).

In April 2018, the European Union (EU) signed a Declaration on Cooperation on Artificial Intelligence (AI) that emphasized the commitment of European Union states towards boosting Europe's technology and industrial capacity in AI. Linking AI technology developments with Health was an inherent step to be done, as shown by the innovation calls launched by H2020 https://cordis.europa.eu/programme/id/H2020_ICT-11-2018-2019: HPC and Big Data enabled Large-scale Test-beds and Applications.

In 2020, the Next Generation EU recovery instrument and the Annual Sustainable Growth Strategy 2021 were launched by the European Commission to fight the effects of the COVID-19 pandemic, providing €750 billion funding. The EU4Health 2021-2027 programme, EU's response to COVID-19, is one the main pillars of this strategy and has a budget of €9.4 billion, which amongst other objectives, will strengthen health systems so that they can face epidemics as well as long-term challenges by stimulating their digital transformation. In this context, Big Data, Artificial Intelligence (AI) and supercomputers, with their analytical power, are major assets in detecting patterns in the spread of the virus or potential treatments.

The European Commission will continue investing in the use of Artificial Intelligence to speed up the diagnosis of COVID-19 and improve future treatment of patients.

Not only public health authorities or institutions, but also private health care providers are incorporating AI technologies in their activities, as a way to gain competitiveness, optimize diagnostic processes, and improve treatment monitoring and lately, to join efforts to fight against COVID-19 pandemics.

12.3 DeepHealth concept

Two of the most known areas of healthcare that can benefit from the advances in Artificial Intelligence are medical imaging and electronic Health Record, as it is highlighted in “AI in Healthcare Whitepaper” (BDVA Task Force7- Sub-group Healthcare- November 2020). The main goal of the DeepHealth project is to put HPC computing power at the service of biomedical applications, to apply Deep Learning and Computer Vision techniques on large and complex biomedical datasets to support a new and more efficient way of diagnosis and to generate insights into complex diseases in a scalable and efficient way.

The DeepHealth concept is based on the scenario, depicted in Figure 1, where processing big quantities of images is needed for diagnosis.

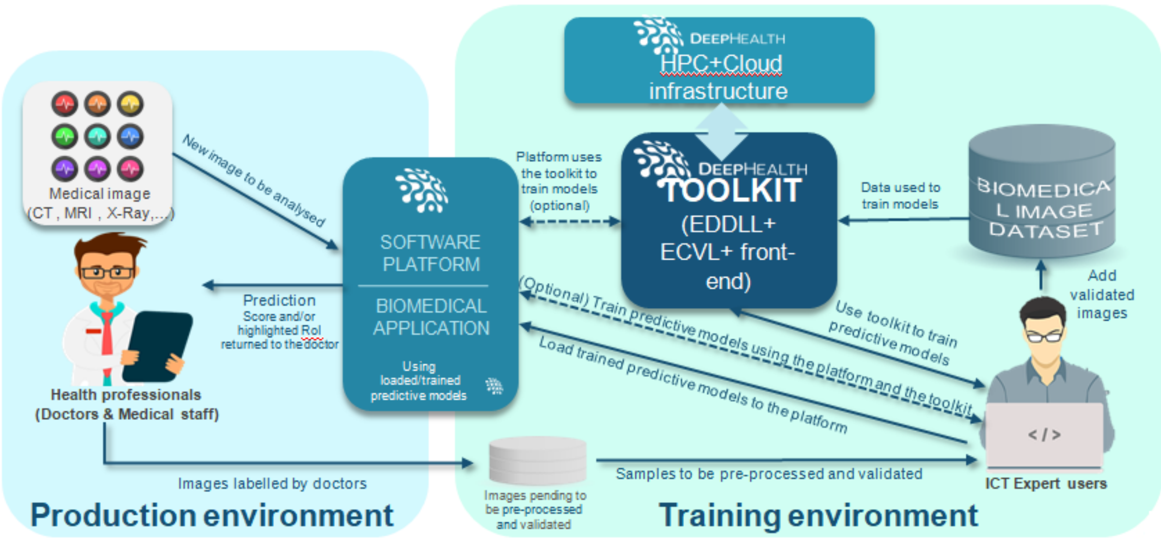

Figure 1 DeepHealth concept

Health Professionals (Doctors and medical staff) are End Users – they are experts in medical areas and diseases.

ICT Experts involved in project will provide medical users with the software applications necessary for obtaining results according to their needs and requirements.

The training environment represents the context where ICT experts work with datasets of images for training predictive models.

In the production environment the medical personnel insert images coming from scanning sessions into a software platform or biomedical application that uses predictive models to get clues that can help them to make decisions during diagnosis. Doctors have the knowledge to label and annotate images, define objectives and provide related metadata. ICT staff is in charge of processing the labelled and annotated images, organizing the datasets, performing image transformations when required, training the predictive models and loading such models into software platforms once tested & validated.

To perform all these operations, ICT staff will use the DeepHealth Toolkit and the DeepHealth HPC & Cloud Infrastructure, as it can be seen in Figure 1.

As it was described also in Chapter 11, the DeepHealth toolkit is a general-purpose deep learning framework, including image processing and computer vision functionalities, enabled to exploit HPC and cloud infrastructures for running parallel/distributed training and inference processes.

The core of the toolkit consists of two libraries, namely the European Distributed Deep Learning Library (EDDL) and the European Computer Vision Library (ECVL) that are ready to be integrated in any software application. Additionally, the DeepHealth toolkit also has two complementary software components, the back end and the front end. The back end exposes a RESTful API to allow the use of all the functionalities provided by the two libraries. The front end makes accessible the functionalities of the two libraries via a web-based Graphical User Interface.

The DeepHealth toolkit incorporates the most advanced parallel programming models to exploit the parallel performance capabilities of HPC and cloud infrastructures, featuring different acceleration technologies such as symmetric multi-processors (SMPs), graphics processing units (GPUs) and field-programmable gate arrays (FPGAs), as shown in Chapter 11.

Moreover, the toolkit provides functionalities to be used for both training and inference, addressing the complexity of the different available computational resources and target architectures at both the training and inference stages.

This process is transparent to doctors; they just provide images to the system and get predictions such as indicators, biomarkers or the semantic segmentation of an image for identifying tissues, bones, nerves, blood vessels, etc.

To demonstrate and validate the concept of the project, the DeepHealth open-source toolkit and the HPC & Cloud infrastructure are tested and validated in real environments thanks to the fifteen pilot use cases (test beds) and on seven biomedical commercial or research platforms, raising the innovation potential of European companies.

12.4 DeepHealth Use Cases

This section will summarize the main aspects that are common to all DeepHealth uses cases, i.e. the testing and validation workflow and KPIs.

A typical Use case in the DeepHealth project is based on a medical imaging data set, which is trained and tested on DeepHealth Toolkit and one or more commercial health platforms, taking advantage of Hybrid and Heterogeneous HPC + Big Data clusters. First, data scientists and members of the team pre-process (labelling, annotation, anonymization) and prepare the dataset by splitting it into three subsets, namely, training, validation and testing subsets. Next the development team designs several artificial neural networks and launches the training processes on HPC and cloud architectures by means of the runtime of the toolkit adapted to HPC frameworks like the ones described in Chapter 11. The team evaluates the models using the validation subset, and redesigns some models if necessary. Sometimes, the team should come back to consider the dataset itself with the knowledge learned in previous iterations. The model that gets the best accuracy using the testing subset is selected; then computer scientists, members of the same team, configure an instance of the application with the best model and deploy the solution in a production environment.

The most important Key Performances Indicators that will be validated and tested in DeepHealth pilots are the time-to-model-in-production (**ttmip**), time-of-pre-processing-images (**toppi**) and time-of-training-models (**totm**).

- **Time of pre-processing images** (**toppi**). This KPI is the sum of T1 and T2, where T1 is the time a data scientist (or ICT expert) needs to design the pipeline of transformations to be applied to the images and the time he/she needs to implement that pipeline of transformations using the ECVL. T2 is the time to run the pipeline over all the images in the dataset.
- **Time of training models (totm).** This KPI is calculated by measuring the execution time of the training procedure.
- **Time to model in production (ttmip).** This KPI is the sum of various variables, the definitive model used in production will commonly be the result of choosing one among several tested models, so the time of designing and testing each model must be taken into account (**totm** is part of the total time necessary for testing each single model). Additionally, if each model needs a particular pipeline of image operations, the value of **toppi** for each defined image-pre-processing procedure must be included as part of the **ttmip**.

Other KPIs that will measure the performance of the algorithms for training DNNs on distributed architectures are speedup and efficiency of parallelism.

Each use case also has specific KPIs to assess that the obtained predictive models provide accuracies equivalent to those obtained with the same DNN topologies when the used models are trained using other toolkits. These KPIs are accuracy, precision, recall and F1-score for classification tasks and intersection over union (IoU) in semantic segmentation tasks. IoU can be computed in average for all the classes to be detected, and individually for each one in order to know in more detail which tissue types or skin spots or tumour types are more difficult to be detected by the DNN-based models.

The fifteen use cases validate the DeepHealth Toolkit performances in the following medical fields (1) neurological diseases, (2) tumour detection and early cancer prediction, and (3) digital analysis of brain pathologies and automated image annotation, exemplifying how the Toolkit can create specific biomedical applications. At the end of year 2020 a new use case was added with a COVID-19+[1] related dataset released by the partner FISABIO.

12.5 Use of HPC& Cloud in medical pilots

As commented in previous chapters, medical datasets are in a constant growth and they are needed as large datasets as possible to train robust and accurate enough models based on CNNs. However, due to the fact that the process of training neural networks requires to iterate over all the samples of the training subset, and the number of iterations / epochs ranges from 100 to 1000 or more, depending on the use case and the architecture, the use of HPC and cloud computing infrastructures are becoming indispensable to distribute the workload by splitting the training subset. In this way each worker node in the computing infrastructure is in charge of performing the back propagation algorithm on a data partition. Chapter 11 provides details on how to distribute the workload on different computing architectures supported by the DeepHealth toolkit. As Figure 1 illustrates, leveraging all the computer power offered by hybrid HPC + Cloud infrastructures equipped with hardware accelerators and many-core CPUs is completely transparent to medical personnel, and ICT experts do not

[1] https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/

need to have a deep knowledge of programming on distributed environments. Thanks to the DeepHealth toolkit, ICT experts can run the training processes on hybrid HPC + Cloud infrastructures, test & validate the trained models, and, finally, update the models in production with the new ones trained with updated datasets after including recently acquired medical images.

Five out of the fifteen use cases of the DeepHealth project are described in this chapter to show the jointly use of ECVL and EDDL. These five uses cases are:

- UC2 UNITOPatho, based on whole-slide colorectal images obtained from colonoscopies;
- UC3 UNITOBrain, based on CT scans of the brain;
- UC4 Chest, based on CT scans of lungs;
- UC5 UNITO Deep Image Annotation, based on Xray chest images;
- UC12 Skin Cancer Melanoma Detection, based on dermoscopic images.

12.5.1. UC2 - UNITOPatho

The expansion of cancer screening programs and the demand of colonoscopy surveillance routines are leading the gastrointestinal histopathology to grow.

Predictive signals of possible gastrointestinal cancer development are colorectal polyps, which are pre-cancerous lesions located in the lining of the colon.[2]

Colorectal tissue samples are collected by biopsies and colonoscopies. Gastrointestinal pathologists examine them to find signs that predict tissue neoplastic process and invasive carcinomas.

[2] Roisin Bevan and Matthew D Rutter,"Colorectal cancer screening—who, how, and when?,"Clinical endoscopy, vol.51, no. 1, pp. 37, 2018

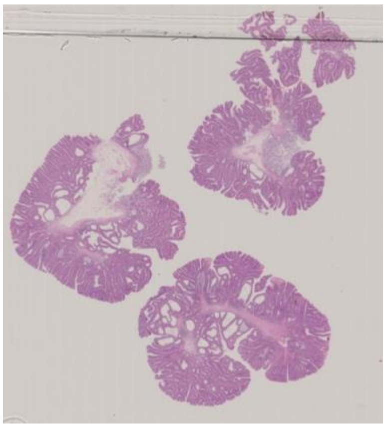

Figure 2 An example of Whole-Slide Image containing Tubulo-villous adenoma tissue

In this use case, we focus our effort to develop a neural network-based pipeline to automatically diagnose colorectal cancers from Whole-Slide Images (WSI).

WSI are super high-resolution images retrieved by scanning biopsies material, more specifically Hematoxylin and Eosin-stained slices of tissue, from patients undergoing cancer screening. These images can reach more than 100.000×100.000 pixels in their size. The WSI corresponds to the whole specimen collected from the patient: it is provided with a label (the diagnosis) and the annotation on the subset of tissue taken into consideration by the pathologist to elaborate the diagnosis (the so-called diagnostic tissue).

Given the nature of the problem, learning from huge images is a hard task which requires HPC infrastructure in order to successfully perform training and, most importantly, inference at diagnosis time. The DeepHealth toolkit is also designed to efficiently handle large-scale images and to train deep-learning models efficiently.

The experts collected six different types (or classes) of tissue: Normal tissue, Hyperplastic polyp, Tubular adenoma with high and low-grade dysplasia, Tubulo-villous adenoma with high and low-grade dysplasia.

A first version of this collection is the open-access dataset UNITOPATHO.[3] It consists of

[3] Luca Bertero, Carlo Alberto Barbano, Daniele Perlo, Enzo Tartaglione, Paola Cassoni, Marco Grangetto, Attilio Fiandrotti, Alessandro Gambella, Luca Cavallo. (2021). UNITOPATHO. IEEE Dataport. https://dx.doi.org/10.21227/9fsv-tm25

9536 hematoxylin and eosin-stained patches extracted from 292 whole-slide images.

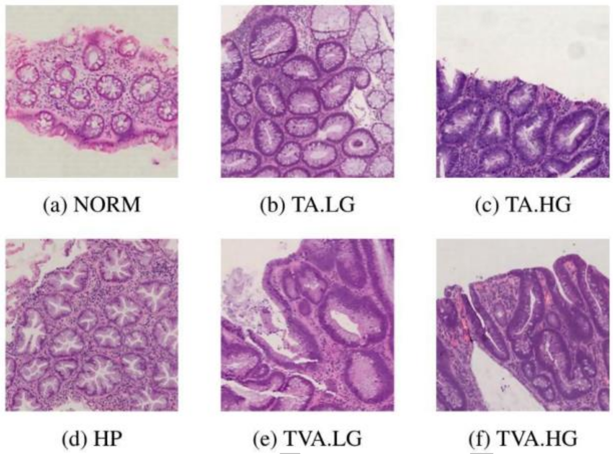

Figure 3 UniToPatho logo, followed by a collection of type-representative tissue samples

In order to solve the classification task, we propose the use of a residual neural network model (ResNet). The neural network model predicts the class for each tissue patch from the WSI and, finally, a label for the WSI itself[4,5]. This UC relies on the University of Torino's OpenDeepHealth (ODH) platform, which implements a hybrid HPC/cloud infrastructure to effectively support the training and inference of AI models. As detailed in chapter 11, ODH integrates the DeepHealth toolkit via Docker containers, both on bare metal and in the multitenant Kubernetes cluster and it has been specifically designed to support AI application on critical data, such as biomedical images.

12.5.2 UC3 - UNITOBrain

The occlusion of a cerebral vessel causes a sudden decrease of the blood perfusion of the corresponding vascular territory. Identifying such an occlusion in a fast and reliable way is

[4] Perlo, Daniele, et al. "Dysplasia grading of colorectal polyps through CNN analysis of WSI." *arXiv preprint arXiv:2102.05498* (2021).

[5] Barbano, Carlo Alberto, et al. "UniToPatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading." *arXiv preprint arXiv:2101.09991* (2021).

critical under emergency scenarios.[6] The so-called CT perfusion, with a sample time of roughly1 Hz, measures the passage of a contrast media bolus into the brain, on a pixel-by-pixel basis. Serial low-dose scans are acquired; time-density curves, corresponding to the contrast media passage in brain tissue, are calculated; parametric maps are calculated. The most relevant parameters used in clinical practice are Cerebral Blood Volume and Cerebral Blood Flow, (CBF and CBV).[7]

Given the nature of the task, generating the aforementioned parametric masks in the least time possible is crucial. Towards this end, HPC infrastructure offers parallel computation capabilities which can be successfully exploited through the use of the DeepHealth toolkit in order to train and to infer from an AI-based model. Data from 115 patients has been collected: a subset of 100 patients has been used to train the deep model, while the remaining 15 were held to validate the results.

In order to generate ground-truth (GT) maps, we relied on a state-of-the-art deconvolution-based algorithm.[8] The validation step has been carried out evaluating concordance between more expert medical evaluators among the segmented lesions using the GT maps and those produced by the artificial neural network model.

As an artificial neural network, we have taken inspiration from the state-of-the-art U-Net architecture.[9] Since the model has been originally thought to segment medical images, and in our case we aim at generating parametric images, we introduce some changes to the standard model, like reducing the granularity of the convolutional operations and using average pool

[6] Donahue J, Wintermark M. Perfusion CT and acute stroke imaging: Foundations, applications, and literature review. Journal of Neuroradiology 2015. https://doi.org/10.1016/j.neurad.2014.11.003

[7] Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. New England Journal of Medicine 2018. https://doi.org/10.1056/NEJMoa1713973.

[8] Bennink E, Oosterbroek J, Kudo K, Viergever MA, Velthuis BK, de Jong HWAM. Fast nonlinear regression method for CT brain perfusion analysis. Journal of Medical Imaging 2016. https://doi.org/10.1117/1.jmi.3.2.026003

[9] Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nature Methods 2019. https://doi.org/10.1038/s41592-018-0261-2.

layers instead of maxpool.

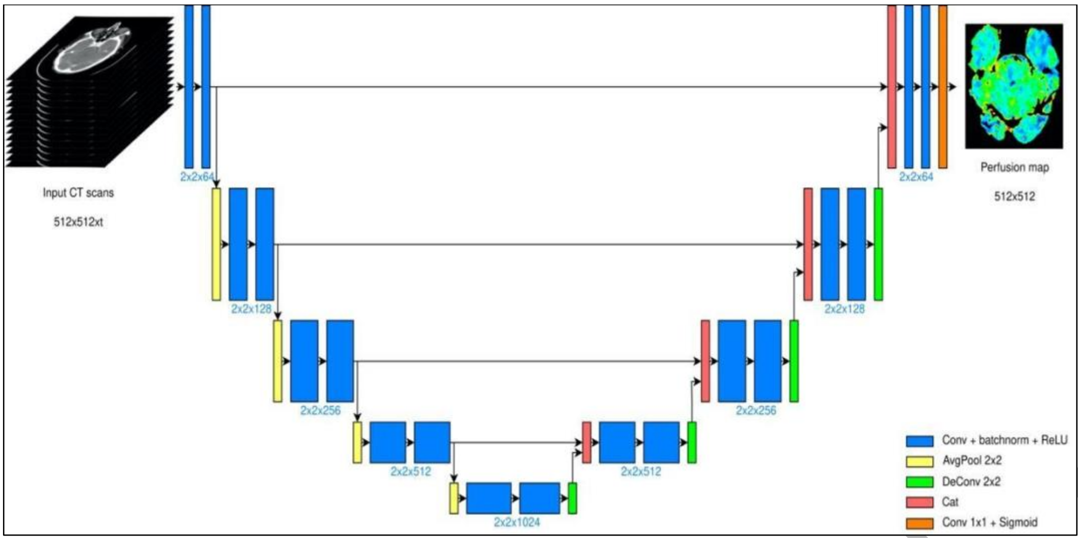

Figure 4 Artificial neural network model deployed for UC3

Contrarily to the other state-of-the-art approaches, no extra information (like the arterial input function) has been provided to the U-Net model, favouring its use in an emergency scenario where the time to obtain these maps is critical.

Overall, on the generated maps, we achieve an average Dice score on the lesions above 0.70, resulting in a generation of good quality perfusion maps from the U-Net model. Inter-rater concordance has been measured, finding a very strong correlation between lesion volumes of CNN maps and GT maps, achieving above 0.98 Pearson correlation.[10] This UC leverage on the HPC features of ODH platform, previously described.

12.5.3 UC4 - Chest

Lung nodules are small focal lesions in the lung parenchyma can be solitary or multiple and in many cases are accidentally found in CT scans. Their identification is time consuming in the current clinical activity for the radiologist and, since these small lesions are difficult to spot, patients often need to perform follow-up CT scans in order to assess their

[10] Gava, Umberto A., et al. "Neural Network-derived perfusion maps: a Model-free approach to computed tomography perfusion in patients with acute ischemic stroke." *arXiv preprint arXiv:2101.05992* (2021).

benignity/malignancy, resulting in increased radiation exposure and anxiety for the patient and increased work amount for doctors. Lung nodules are quite common incidental findings in CT scans and can be defined as small focal lesions (ranging from 5 to 30 mm) that can be solitary or multiple. Deep learning models outperform traditional computer vision techniques in various tasks. Typically, features of the input data are hand-crafted, but deep-learning features are learned in an end-to-end fashion. Convolutional Neural Networks (CNNs) are one of the most popular models that are also employed in medical imaging.[11] More specifically, the U-Net model is used to identify lung nodules in CT scans. The dataset used to train and evaluate the network is provided by Citta' della Salute e della Scienza di Torino.

	Patients	Images
Train	247	13589
Validation	61	1699
Test	109	1708

In order to get preliminary results, splits have been created with only images with a ground truth mask, such that the training set contains 80% of images, validation 10% and test 10% CT scans are in DICOM format.

CT-scans are very large images: every exam may contain more than 200 acquisitions which need to be properly processed. This use case exploits HPC by means of the ODH platform to reduce the computation time. Handling these image format and training the deep-learning model to perform the segmentation task efficiently is crucial for the success of this task: the DeepHealth toolkit offers efficient the capability of training segmentation models on medical images in an efficient way and is therefore crucial towards the success of this use-case.

[11] G. Nam et al., "Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs, "Radiology, vol. 290,no. 1, pp. 218–228, 2019.

The model currently reaches an Intersection Over Union of 0.62 with only rotations as data augmentation.

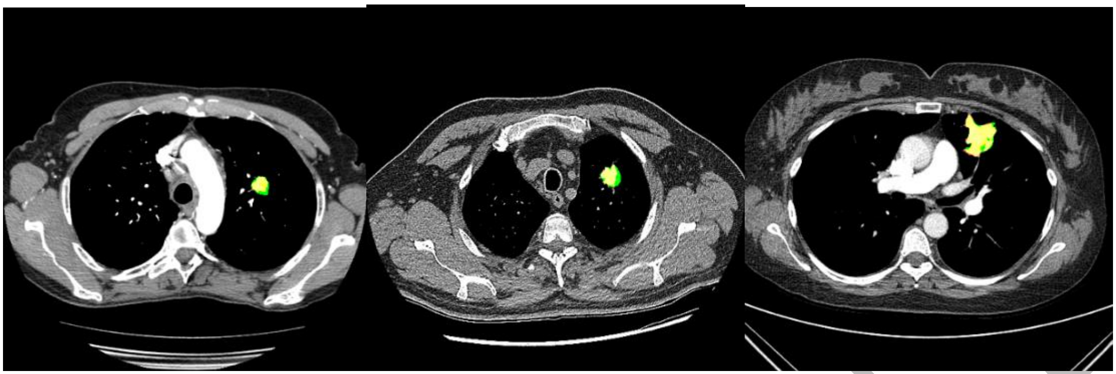

Figure 5Segmentations from the U-Net model in green, ground-truth in yellow

12.5.4 UC5- UNITO Deep Image Annotation

Statistics prepared by the European Union clearly show that the number of imaging devices and, consequently, the number of daily scans is steeply increasing all over Europe[12]. The large amount of digital images collected daily by the national health systems or even by single hospitals poses new challenges and a huge burden on the clinical community. Indeed, radiologists need to inspect every single image and write detailed reports; often taking into account also information stored in Electronic Health Records (EHR). Such a heavy workload may easily lead the operators to make mistakes not so much in providing the correct diagnosis but rather in the repetitive and tedious data entry and reporting procedures. Information Technology (IT) and, specifically, the innovative solutions based on Artificial Intelligence (AI) implemented in several use cases of the DeepHealth project might help also in reducing the mistakes caused by habituation and fatigue.

[12] EU Healthcare resource statistics, http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthcare_resource_statistics_-_technical_resources_and_medical_technology&oldid=280129

Figure 6Two examinations composed of two views of the chest. Upper row: normal examination and related report with the sections used in UC5. Lower row: anomalous examination and fragment of the related report with the diagnosis and the MeSH/MTI terms (labelled as "automatic")

UC5 "Deep Image Annotation" sets its main goal to the automatic reporting of medical images, i.e., given one or more input images, UC5 aims at generating automatically a description of the image content using natural language.

The multimodal processing in UC5 is based on a combination of two different artificial neural networks (ANN) set in a processing pipeline[13]. In the first stage, a convolutional neural network (CNN) classifies and encodes the input images; in the second stage, a recurrent neural network (RNN) generates sentences starting from the encoding coming from the CNN. UC5 uses the public and anonymized dataset named "Indiana University chest X-ray Collection"[14] containing 7470 X-ray images and 3955 semi-structured reports (see Fig. 6). Each report, in XML format, is paired with two views of the chest and contains multiple textual annotations, including sections with the "indication", the "findings" and the "impression", that correspond to the actual report written by the radiologists. The reports

[13] B. Jing, P. Xie, E. P. Xing. "On the Automatic Generation of Medical Imaging Reports", Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, pages 2577–2586 Melbourne, Australia, July 15 - 20, 2018.

[14] D. Demner-Fushman et al. "Preparing a collection of radiology examinations for distribution and retrieval." *Journal of the American Medical Informatics Association* 23.2, pages 304-310, 2016.

contain also MeSH and MTI (respectively, Medical Subject Headings and Medical Text Indexer, both by the National Library of Medicine, USA) terms (or tags).

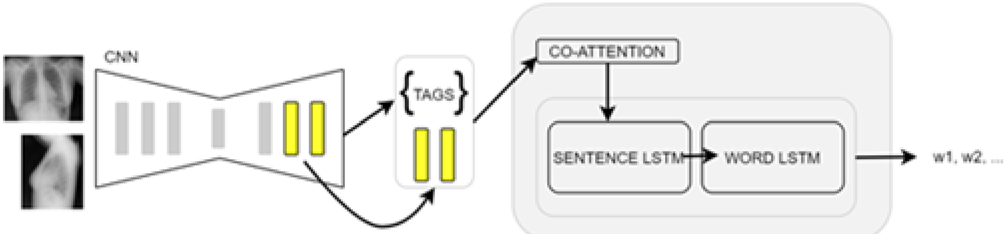

Figure 7Neural network architectures used in UC5. The images are processed by a CNN, and then their classification and embeddings are passed to a MLP (Co-attention), that feeds its output to the cascaded LSTM, the first generating topics of the sentences, the second generating words according to the topic

The two networks are trained independently. First, the CNN is trained to assign one or more tags, extracted from the reports, to the input images. We use a VGG-19[15] in the CNN module, trained in a multi-label fashion: its training set is composed of images with an encoding of tags as target values. The trained CNN is then used to build the training dataset for the RNN: for each image in the training set, we first extract the hidden representations (encoding) from the two final convolutional layers (located before the output layer), then we append the concatenation of those two layers to the representation of each word in the sections "findings" and "impression". The RNN module is then trained to generate the same sequence of words it receives as input. It is composed of two Long-Short Term Memory networks[16] organized in a hierarchical way: the first LSTM, receiving the semantic features by the CNN module (actually, by an intermediate co-attention feed forward module), generates topic vectors for sentences (hereafter, topic or sentence LSTM) that are fed to the second LSTM, responsible of generating sequences of words (hereafter, word LSTM). The sentence LSTM

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[16] S. Hochreiter and J. Schmidhuber. "Long short-term memory." *Neural computation* 9.8, pages 1735-1780, 1997.

receives a co-attention vector[17] that assigns different weights to parts of the semantic features extracted by the CNN, mimicking the visual exploration made by an expert radiologist. The number of topic vectors generated by the first LSTM that corresponds to the number of sentences in the final text is controlled by a multilayer perceptron, trained at the same time as the two LSTMs. We do not provide here the equations or the full details describing the system behaviour that can be found in the original paper[18].

The system is evaluated automatically using the Bilingual Evaluation Understudy (BLEU) score. Furthermore, we plan to make a "crowd"-based evaluation by letting humans grade the quality of the generated texts with respect to the ground truth. However, the evaluation is an open issue: automatic metrics are not able to cope with the specific "jargon" (e.g., sentences like "no anomalies detected" and "anomalies detected" have opposite meanings but receive a high similarity score) and it is not easy to enrol many domain experts in a crowd-based evaluation.

Based on the preliminary results, UC5 will likely not be able to produce reports as reliable as human radiologists do (as expected). Nevertheless, UC5 can still have a positive and measurable impact in the daily clinical routine, even if limited, for example, in speeding up the reporting activity while potentially reducing the number of errors as well (for example, mistakes in copying and pasting from previously written reports[19]). UC5 ``Deep Image Annotation" offers several interesting opportunities. It is computationally challenging, thus providing a good test bed for the two ECVL and EDDLL libraries under development in DeepHealth. If well engineered, it can be a valuable support tool in current commercial

[17] Jing et al. On the automatic, 2018.
[18] Idem.
[19] A.Y. Tsou et al. "Safe practices for copy and paste in the EHR: systematic review, recommendations, and novel model for health IT collaboration." *Applied clinical informatics* 8.1 2017.

systems. And, with more data becoming available, the reports, that it will be able to generate, will represent a precious help for experienced or junior radiologists.

12.5.5. UC12 –Skin cancer melanoma detection

Skin cancer represents a major public health issue, being the most common form of human cancer worldwide with an increasing trend [20]. In the last decades, many efforts have been given to improve skin cancer treatments; however, the early detection remains a key factor in preventing cancer progression to advanced stages and ensuring a lower mortality rate[21].
For epithelial skin cancer, such as basal cell carcinoma and squamous cell carcinoma, delayed diagnosis is mainly responsible for a larger and more disfiguring surgery, which could also affect relevant functional structures (*e.g.* mouth, ears, eyelids, and nose). For melanoma, a delayed diagnosis may mean death due to potential tumour aggressiveness. Early diagnosis represents the ideal and cheap solution to fight against the skin tumour consequences.
Dermoscopy is a form of in-vivo skin surface microscopy performed using high quality magnifying lenses and a powerful light source to mitigate the surface reflection of the skin, to enhance the visibility of the pigmentation of the lesion. This technique is broadly employed by dermatologists since it allows for a fast diagnosis, and significantly increased the diagnosis accuracy, sensitivity, and specificity with respect to the naked eye examination. However, this kind of non-invasive imaging approaches requires the eye of expert clinicians to successfully diagnose skin cancer. Therefore, many efforts have been given towards the development of automatic tools for supporting and training physicians in the analysis of

[20] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394-424.

[21] Rigel, D. S., Russak, J., & Friedman, R. (2010). The evolution of melanoma diagnosis: 25 years beyond the ABCDs. *CA: A Cancer Journal for Clinicians*, *60*(5), 301-316.

dermoscopic images[22]. Due to its outstanding results in many research areas including image understanding and image classification[23] [24], deep learning has become the main option for analyzing medical images in several field, including skin lesion classification[25] [26] [27]. Unfortunately, these algorithms require a huge amount of annotated data to ensure the correct learning process. When dealing with medical imaging, collecting, and annotating data can be cumbersome and expensive. This is mainly related to the nature of the data and to the need for well-trained expert technicians. Moreover, such kind of algorithms requires a significant amount of (expensive) computational resources, which are often inaccessible to medical personnel.

As already introduced in the previous Sections, the main goal of the DeepHealth project is compensating the second issue, boosting the productivity of data scientists operating in the medical field by providing a unified framework for the distributed training of neural networks, which is able to leverage hybrid HPC and cloud environments in a transparent way for the user, without requiring a deep understanding of DNNs and distributed high-performance computing.

[22] Allegretti, S., Bolelli, F., Pollastri, F., Longhitano, S., Pellacani, G., & Grana, C. (2020). Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*.

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[24] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105-6114). PMLR.

[25] Zhang, J., Xie, Y., Xia, Y., & Shen, C. (2019). Attention residual learning for skin lesion classification. *IEEE Transactions on Medical Imaging*, *38*(9), 2092-2103.

[26] Pollastri, F., Parreño, M., Maroñas, J., Bolelli, F., Paredes, R., Ramos, D., & Grana, C. (2021). A Deep Analysis on High Resolution Dermoscopic Image Classification. *IET Computer Vision.*

[27] Canalini, L., Pollastri, F., Bolelli, F., Cancilla, M., Allegretti, S., & Grana, C. (2019, September). Skin lesion segmentation ensemble with diverse training strategies. *In International Conference on Computer Analysis of Images and Patterns (pp. 89-101).* Springer, Cham.

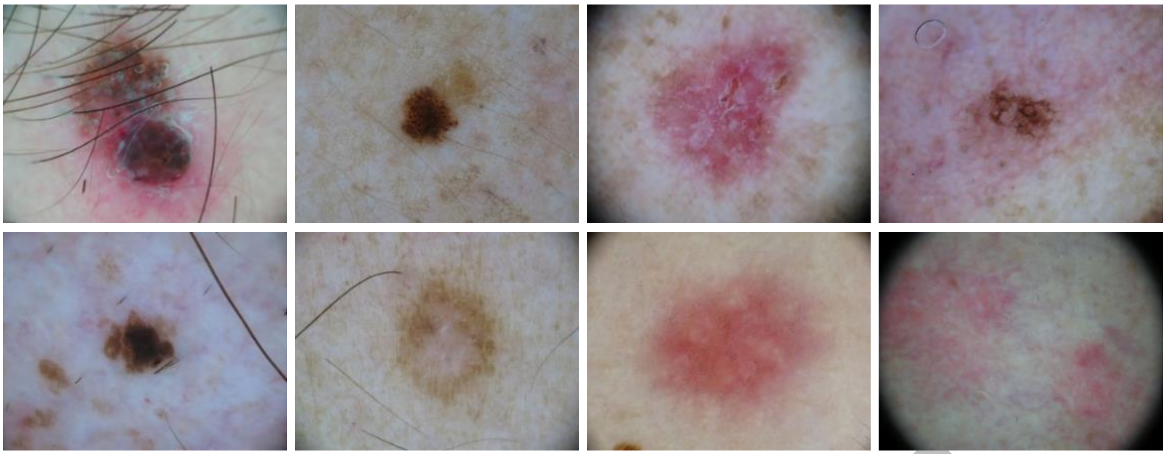

Figure 8 Samples of internal dataset. From upper left to right: Melanoma - MEL, Melanocytic Nevus - NV, Basal Cell Carcinoma - BCC, Actinic Keratosis - AK, Benign Keratosis - BKL, Dermatofibroma - DF, Vascular Lesion - VASC, Squamous Cell Carcinoma – SCC.

On the other hand, UC12 has the main goal of collecting dermoscopic images and design advanced deep learning algorithms for the segmentation and classification of skin lesion, providing clinicians with Computer-Aided Diagnosis (CAD) systems for the automated melanoma recognition. Indeed, segmentation of images and extraction of features can lead to a rapid and automatic identification of diagnostic clues which can facilitate image interpretation and diffusion of technologies among other doctors.

In this sense, the UC12 combines and exploits existing publicly available datasets with a huge internal one, completed of clinical, dermoscopy and confocal microscopy images, annotated with conclusive diagnosis (histologic or clinically confirmed), and relevant patient’s data. Samples of such a kind of images are provided in Fig. 8.

This dataset consists of 25,849 dermoscopy images, collected between 2003 and 2019 using several distinct acquisition tools. The dataset presents a different category distribution compared to the public datasets, with a higher percentage of melanoma cases. Visual artefacts which could be considered source of biases, such as rulers, ink markings/staining, and coloured patches are almost completely absent.

CHAPTER 12

For what concerns public datasets, the International Skin Imaging Collaboration (ISIC) began to aggregate a large-scale collection of dermoscopic skin lesion images in 2016[28]. The 2019 version of the ISIC archive contains a total amount of 25,331 labelled dermoscopic images belonging to nine different classes, which represent eight types of skin lesion plus an additional category named *none of the others*, which contains samples of different natures that do not belong to any of the other eight classes.

Differently from ISIC 2019, the 2020 dataset is focused on a binary classification problem. In this case images are divided in only two classes: melanoma or non-melanoma. Moreover, this set of dermoscopic images contains patient-level contextual information, providing for each image an identifier which allows lesions from the same patient to be mapped to one another. This additional knowledge is frequently used by clinicians to diagnose melanoma and is especially useful in ruling out false positives in patients with many atypical nevi.

This new dataset is composed of 33,126 images and collected from 2,056 patients (21% of them with at least one melanoma, 79% with zero melanomas) gathered from 1998 to 2020, with an average of 16 lesions per patient.

All three datasets provide metadata such as sex, age, site of the lesion, and number of patient lesions.

Image analysis of this use-case focus on the extraction of key parameters (such as pagetoid spreading, atypical cells at the junction, atypical melanocytic nests, alteration of the architecture for melanoma, identification of tumour islands and cords for basal cell carcinoma, identification, and quantification of dyskeratosis for squamous cell carcinoma) useful to obtain an accurate description of histopathologic background from confocal images. This information will be correlated with parameters obtained from dermoscopic image

[28] Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018, April). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 168-172). IEEE.

analysis and integrated with other relevant clinical information in order to provide an accurate diagnosis and a decisional protocol, able to reproduce the analytical process of differential diagnosis of doctors with the highest expertise in the field.

For what concerns classification, this use-case implements an ensemble of state-of-the-art architectures, combined with several techniques for data augmentation, such as random vertical/horizontal flips, random rotation, additive Poisson noise, and dropout. All of these techniques have been implemented by means of ECVL and EDDL libraries, developed within the project.

The ensemble prediction output combines six different neural networks based on EfficientNet-Bx and ResNet152 state-of-the-art architectures. The single models have been trained with different input and batch sizes and using different augmentation strategies. Each network is trained for ~20 epochs on ISIC 2019 and ISIC 2020 using the Cross-Entropy loss and Adam[29] optimizer, with a learning rate of 3e-5. Two fully connected layers process the dataset metadata and then the output is concatenated those of the networks to be assessed as well. The ensemble achieves a 0.94 AUC in the melanoma/non melanoma classification problem on the official test set of ISIC 2020 and 0.87 on the internal dataset.

The second problem issued by the use-case concerns the segmentation of skin lesion images with the aims of producing a segmentation mask for a certain input image (**Error! Unknown switch argument.**). The quality of output segmentation results is, again, ensured by the ensemble of different state-of-the-art architectures, mostly based on DeepLabv3+[30], achieving the outstanding mIoU (mean Intersection over Union) of 0.850.

[29] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[30] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

Figure 9 Example of segmented (red contour) skin lesion images from the ISIC dataset.

The *DeepHealth Service* acts as the link between the DeepHealth libraries implementing the previously described models and the HPC and Cloud systems. Using the *DeepHealth Service*, data scientists do not have to write code for the DeepHealth library APIs or directly manage computing resources, but directly use ECVL and EDDL functionalities through a RESTful web service and, optionally, a web-based GUI. This interface provides clinicians useful tools that can assist physicians with dermatologist-grade decision support.

Additionally, the *Service* provides the ability to design, train and test additional predictive models and to perform pre- and post- processing without writing any code. And allowing evolving and improving the existing use-case whenever new state-of-the-art architectures will be released published in literature.

Instead, the REST interface enables managed service usage scenarios, where a potentially complex and powerful computing infrastructure (e.g., high-performance computing, cloud computing or even heterogeneous hardware) could be transparently used to run deep learning jobs without the users needing to directly interface with it.

A specific instance of the *Service* has been deployed for the UC12 tasks and it has been configured for the asynchronous and distributed training (and test) of aforementioned classification and segmentation neural networks. With this goal, the *Service* spreads the jobs among different cloud nodes which have several GPUs each.

12.6 DeepHealth Value Proposition

As it has been previously described in Section 12.3 DeepHealth concept, DeepHealth is addressed to two kinds of users: ICT expert users who train AI models and healthcare professionals who use the trained models to predict on images.

DeepHealth provides a multi-step value proposition addressing all the potential users of these two categories. Each value proposition is adapted to each of the target users and presented in the format: For - who needs – That- Unlike and, finally, the positioning statement:

- For **Research Institutions** WHO NEED state-of-the-art, open-source DL&CV technologies to advance in their research concerning AI for imaging with a reduced time-to-model THAT allow them to make an optimised use of the latest hardware technologies in a flexible and easy way UNLIKE non-European tools, not adapted to European HPC and hardware accelerators, nor integrated in European platforms, which are less innovative or not open source. DeepHealth Toolkit is positioned as "The European, open-source DL&CV framework, modular and scalable, that streamlines development of models, thanks to an optimised use of state-of-the-art HPC technologies."
- For **Independent expert users and start-ups** WHO NEED technologies to offer highly accurate AI models for medical images in less time, with higher accuracy and better diagnostic functionalities, by having a technological differential in order to expand their business THAT allow them to generate innovative AI models thanks to transparent use of HPC technologies; including specific functions for biomedical images and being a unified easy-to-configure framework, with an easy-to-install back-end and user-friendly front-end UNLIKE non-open source tools which imply an

additional cost; other OSS tools non-optimized for HPC architectures & hardware accelerators; other OSS tools difficult to configure, deploy and orchestrate among them and which do not include an integrated back end and front end; or other OSS tools not specifically designed for biomedical imaging, which may lack specific functionalities. DeepHealth Toolkit is positioned as an "An all-in-one, open-source DL&CV framework specialized for developing innovative medical imaging AI solutions, with higher accuracy and for new use cases thanks to the optimised use of HPC technologies."

- For **Healthcare Technology Vendors** WHO NEED new, easy to integrate open-source tools to expand their offering in order to provide the latest technological innovations, with special focus on improving expert users' productivity and addressing new medical imaging use cases THAT allow them to stay competitive, by reducing the time-to-market of their models and increasing their accuracy thanks to multiplatform frameworks for transparent exploitation of different high performance hardware UNLIKE non-multiplatform OSS tools, difficult to integrate and without associated professional services (consulting, support, etc.); OSS tools which do not accelerate time-to market; Non-innovative OSS tools, which do not include specific methods for distributed computing and adapted middleware or which are not transparent in the use of hardware. DeepHealth Toolkit is positioned as "An open-source DL&CV framework that makes transparent the optimised use of state-of-the-art HPC technologies with the flexibility of a multiplatform solution, increasing expert users' productivity, reducing time-to-market. Specialized for healthcare and validated by fifteen biomedical imaging Use Cases."
- For **Healthcare professionals** WHO NEED support in diagnosis due to growing shortage of imaging specialists, increasing number of patients and to decrease

erroneous diagnosis and healthcare costs THAT allow them to leverage the vast amounts of datasets available and benefit from the high value provided by AI technologies, thanks to highly accurate algorithms validated by clinicians and developed in collaboration with them and with an adapted interface for clinicians' needs UNLIKE non-reliable algorithms, developed with few iterations to decrease development time; AI solutions which cannot address specific use cases due to the high computing capabilities necessary; Long time-to-model AI solutions for urgent use cases, like COVID-19 lung predictive models; or complex software solutions which cannot be easily integrated in the clinical workflow. DeepHealth Solutions are positioned as "Reliable, highly accurate AI-for-imaging solutions, developed in collaboration with expert clinicians and validated by them, for high added-value use cases and integrated in user-friendly platforms adapted for healthcare professional needs."

12.7 Conclusions

In conclusion, the validation of the DeepHealth concept in large scale pilots will support the impact and the benefits the project is expected to have.

Using the Deephealth toolkit and taking advantage of HPC and Cloud architectures will increase the productivity of ICT staff working in the health sector by allowing them to design, train and test significantly more predictive models in the same time period.

This will facilitate the daily work of expert users that are managing large image or other types of datasets; AI systems used in radiology could outperform human experts or aid them by reducing their workload.

Knowledge about diseases and pathologies will be extended by applying the DeepHealth Toolkit. Early diagnosis and improving treatments will be possible, finally impacting the well fare of people and saving direct and indirect healthcare costs. Outcomes of the project will be useful to other sectors: EDDLL is a general-purpose Deep Learning Library and ECVL will be useful for image processing in general.

Other industries can easily adopt the DeepHealth Toolkit, following the trend AI+HPC as a service for an increasing number of applications (Other DL-based applications, Graph-based applications: data-discovery, digital Twins and more…). The project thus contributes to increasing the impact of AI on European society.

It is anticipated that results of the project will be able to support both private companies and public institutions in the Healthcare domain.