UNIVERSITÀ DELLA CALABRIA

UNIVERSITÀ DELLA CALABRIA

Dipartimento di Matematica e Informatica

Dottorato di Ricerca in Matematica e Informatica

XXXIII CICLO

Towards an effective and explainable AI: studies in the biomedical domain

Settore Scientifico Disciplinare INF/01 - INFORMATICA

Coordinatore: Ch.mo Prof. Gianluigi Greco

Greco Gianluigi
02.07.2021 18:17:24
GMT+00:00

Supervisore: Prof. Francesco Calimeri

CALIMERI FRANCESCO
02.07.2021 16:43:41
GMT+00:00

Dottoranda: Dott.ssa Pierangela Bruno

PIERANGELA BRUNO
02.07.2021 16:28:00
UTC

A mia madre.

Acknowledgments

First of all, I would like to thank my supervisor Prof. Francesco Calimeri for his invaluable supervision, continuous support, motivation, and patience. His knowledge inspired me and his advice guided my research.

I would like to thank Prof. Dr. Lena Maier-Hein, Tobias, and the whole CAMI group of the DKFZ in Heidelberg for embracing my research visiting period in Germany and the opportunity to work in a wonderful context. Moreover, special thanks go to Prof. Elena De Momi and Prof. Maria Francesca Spadea for their encouragement and guidance over my Ph.D. years. I truly thank Sara Moccia and Paolo Zaffino for their support and suggestions throughout my research activities.

I would like to thank the Department of Mathematics and Computer Science, and the people working there, for providing me the valuable opportunity of this Ph.D. research and for the kind support, and also the coordinator of the Ph.D. program in Mathematics and Computer Science, Prof. Gianluigi Greco.

I am sincerely thankful to Cinzia, Paola, and Marco for always listening to me and giving me words of encouragement. Also, I would like to thank all my colleagues and friends I met during these three years.

Ringrazio i miei amici per il loro supporto e affetto.
Ringrazio la mia famiglia e Paolo per aver sempre creduto in me e per tutto l'amore dimostratomi sempre e incondizionatamente.

Abstract

Providing accurate diagnoses of diseases and maximizing the effectiveness of treatments requires, in general, complex analyses of many clinical, omics, and imaging data. Making a fruitful use of such data is not straightforward, as they need to be properly handled and processed in order to successfully perform medical diagnosis. This is why Artificial Intelligence (AI) is largely employed in the field. Indeed, in recent years, Machine Learning (ML), and in particular Deep Learning (DL), techniques emerged as powerful tools to perform specific disease detection and classification, thus providing significant support to clinical decisions. They gained a special attention in the scientific community, especially thanks to their ability in analyzing huge amounts of data, recognizing patterns, and discovering non-trivial functional relationships between input and output. However, such approaches suffer, in general, from the lack of proper means for interpreting the choices made by the learned models, especially in the case of DL ones.

This work is based on both a theoretical and methodological study of AI techniques suitable for the biomedical domain; furthermore, we put a specific focus on the practical impact on the application and adaptation of such techniques to relevant domain.

In this work, ML and DL approaches have been studied and proper methods have been developed to support (i) medical imaging diagnostic and computer-assisted surgery via detection, segmentation and classification of vessels and surgical tools in intra-operative images and videos (e.g., cine-angiography), and (ii) data-driven disease classification and prognosis prediction, through a combination of data reduction, data visualization and classification of high-dimensional clinical and omics data, to detect hidden structural properties useful to investigate the progression of the disease. In particular, we focus on defining a novel approach for automated assessment of pathological conditions, identifying latent relationships in different domains and supporting healthcare providers in finding the most appropriate preventive interventions and therapeutic strategies. Furthermore, we propose a study about the analysis of the internal processes performed by the artificial networks during classification tasks, with the aim to provide a AI-based model explainability.

This manuscript is presented in four parts, each focusing on a special aspect of DL techniques and offering different examples of their application in the biomedical

domain.

In the first part we introduce clinical and omics data along with the popular processing methods to improve the analyses; we also provide an overview of the main DL techniques and approaches aimed at performing disease prediction and prevention and at identifying bio-markers via biomedical data and images.

In the second part we describe how we applied DL techniques to perform the segmentation of vessels in the ilio-femoral images. Furthermore, we propose a combination of multi-instance segmentation network and optical flow to solve the multi-instance segmentation and detection tasks in endoscopic images.

In the third part a combination of data reduction and data visualization techniques is proposed for the reduction of clinical and omics data and their visualization into images, with the aim of performing DL-based classification. Furthermore, we present a ML-based approach to develop a risk model for class prediction from high-dimensional gene expression data, for the purpose of identifying a subset of genes that may influence the survival rate of specific patients.

Eventually, in the fourth part we provide a study on the behaviour of AI-based systems during classification tasks, such as image-based disease classification, which is a widely studied topic in the recent years; more in detail, we show how DL-based systems can be studied with the aim of identifying the most relevant elements involved in the training processes and validating the network's decisions, and possibly the clinical treatment and recommendation.

Sommario

Per fornire diagnosi accurate o massimizzare l'efficacia di una terapia bisogna, in genere, analizzare con oculatezza un ingente numero di dati clinici, omici e immagini biomediche. Tuttavia, tali dati non sono di facile lettura; anzi, necessitano di una adeguata gestione e di complesse elaborazioni. In tali contesti, l'Intelligenza Artificiale, il cui uso è ormai ampiamente diffuso, può giocare un ruolo cruciale. Negli ultimi anni, infatti, le tecniche di Machine Learning (ML), e in particolare di Deep Learning (DL), si sono affermate come strumenti indispensabili per eseguire il rilevamento e la classificazione di malattie specifiche, fornendo così un supporto importantissimo al personale medico. Con il tempo queste tecniche hanno generato un'attenzione crescente all'interno della comunità scientifica, soprattutto grazie alla loro capacità di analizzare un'enorme quantità di dati, riconoscere specifici modelli e scoprire relazioni funzionali non banali tra input e output. Tuttavia, tali approcci soffrono della mancanza di strumenti adeguati per interpretare le scelte operate dai modelli appresi, soprattutto nel caso di quelli basati su DL.

Questo lavoro si basa su uno studio teorico e metodologico di tecniche di Intelligenza Artificiale adatte al contesto biomedico; inoltre, ci siamo concentrati in modo particolare sull'impatto pratico dell'applicazione e dell'adattamento di tali tecniche a domini di particolare rilevanza.

Al centro di questo lavoro c'è lo studio di approcci di ML e DL e lo sviluppo di metodi adeguati per supportare (i) la diagnostica per immagini mediche e la chirurgia assistita da computer tramite il rilevamento, la segmentazione e la classificazione di vasi sanguigni e strumenti chirurgici in immagini e video intraoperatori (ad es. cine-angiografia) e (ii) la classificazione di malattie e la previsione di prognosi basata sui dati, attraverso una combinazione di riduzione, visualizzazione e classificazione di dati clinici e omici ad alta dimensionalità, al fine di rilevare proprietà strutturali nascoste e potenzialmente utili per predire la progressione della malattia. In particolare, ci siamo concentrati sulla definizione di nuovi approcci per automatizzare la valutazione della condizione patologica, e sulla identificazione di relazioni latenti in diversi domini applicativi in ambito medico, al fine di supportare gli operatori nella ricerca di interventi preventivi e di strategie terapeutiche appropriate. Infine, proponiamo un'analisi volta alla comprensione dei processi interni eseguiti dalle reti neurali durante le attività di classificazione, con lo scopo di fornire una spiegazione

dei modelli basati sull'intelligenza artificiale.

Questo elaborato è strutturato in quattro parti, ciascuna incentrata su un aspetto specifico delle tecniche di DL e su diversi esempi della loro applicazione nel campo biomedico.

Nella prima parte introduciamo i dati clinici e omici insieme ai più comuni metodi di elaborazione usati per migliorare le analisi. Forniamo inoltre una panoramica delle principali tecniche e dei principali approcci di DL finalizzati alla previsione e prevenzione delle malattie e all'identificazione di biomarcatori tramite dati e immagini biomedici.

La seconda parte, invece, è incentrata sulla descrizione dell'applicazione di tecniche di DL utili ad eseguire la segmentazione dei vasi sanguigni in immagini ileofemorali. Inoltre, propone la combinazione di una rete neurale per la segmentazione di oggetti e tecniche basate sull'*optical flow*, per eseguire la segmentazione e il rilevamento di istanze in immagini endoscopiche.

Nella terza parte proponiamo la combinazione di tecniche per la riduzione di dati clinici e omici e la loro visualizzazione in immagini, con l'obiettivo di eseguire una classificazione basata su DL. Presentiamo, inoltre, un approccio per lo sviluppo di um nodello di previsione della classe di rischio, basato su tecniche di ML e sull'uso di dati di espressione genica, allo scopo di identificare un sottoinsieme di geni che possono influenzare il tasso di sopravvivenza di pazienti specifici.

Infine, nella quarta ed ultima parte ci focalizziamo sullo studio di un argomento ampiamente discusso negli ultimi anni: l'*explainability* dei sistemi basati sull'Intelligenza Artificiale. Ad esempio, presentiamo i risultati di un'analisi per la comprensione delle scelte effettuate da una rete neurale durante la classificazione di malattie basata su immagini. Più in dettaglio, mostriamo come i sistemi basati su DL possono essere studiati con l'obiettivo di identificare gli elementi più rilevanti coinvolti nel processo di formazione, e convalidare così sia le decisioni della rete che, eventualmente, il trattamento clinico e la raccomandazioni mediche.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning
DL	**D**eep **L**earning
ANN	**A**rtificial **N**eural **N**etwork
XAI	**E**xplainable **A**rtificial **I**ntelligence
CAD	**c**omputer-**a**ided **d**iagnosis
EHR	**E**lectronic **H**ealth **R**ecord
ODM	**O**mics **D**ata **M**anagement
CDM	**C**linical **D**ata **M**anagement
NLP	**N**atural **L**anguage **P**rocessing
NGS	**N**ext **G**eneration **S**equencing
MS	**M**ass **S**pectrometry
SVM	**S**upport **V**ector **M**achine
PCA	**P**rincipal **C**omponent **A**nalysis
PAM	**P**rediction **A**nalysis for **M**icroarrays
ClaNC	**Cla**ssification to **N**earest **C**entroids
POS	**P**roportional **O**verlapping **S**core
NSC	**N**earest **S**hrunken **C**entroids
COO	**C**ell-**o**f-**O**rigin
ABC	**A**ctivated **B**-**C**ell-like
GCB	**G**erminal **C**enter **B**-cell-like
DLBCL	**D**iffuse **L**arge **B**-cell **L**ymphoma

AUC	**A**rea **U**nder the **C**urve
LSTM	**L**ong **S**hort-**t**erm **M**emory
MTM	**M**ulti-scale **T**emporal **M**emory
CNN	**C**onvolutional **N**eural **N**etwork
SGD	**S**tochastic **G**radient **D**escent
ReLU	**R**ectified **L**inear **U**nit
RPN	**R**egion **P**roposal **N**etwork
ROI	**R**egion **o**f **I**nterest
PAOD	**P**eripheral **A**rterial **O**cclusive **D**isease
DCE	**D**ice **S**imilarity **C**oefficient
MIS	**M**inimally **I**nvasive **S**urgeries
SDS	**S**urgical **D**ata **S**cience
RNN	**R**ecurrent **N**eural **N**etwork
GradCAM	**Grad**ient-weighted **C**lass **A**ctivation **M**apping

Introduction

A large amount of patient-related information is collected by healthcare operators in their everyday activities, which span over a wide spectrum of medical processes, such as wellness check-ups or examinations at healthcare hospitals or medical offices, just to name a few. For instance, when a patient undergoes a medical examination for the first time, the physician usually creates a patient file including his medical history, current treatments, medications, diagnosis and other relevant information [Evans 1999]. Considering that disease diagnosis is crucial for health condition monitoring, it is natural to envisage that such a large amount of data can be profitably used to guide data-driven disease classification tasks in the quest for early and accurate diagnoses, taking care of the complex interactions among clinical, biological, and pathological variables.

Artificial Neural Networks (ANNs) are a Machine Learning (ML) method inspired to the human brain. Thanks to their ability to identify complex relationship within a data set and detect latent patterns, ANNs have been widely used in many contexts such as image classification, text analysis, speech and facial recognition. Over the last few years, ANN-based approaches have been used in medical applications to provide diagnosis, treatment, and predicting outcomes, with the aim to improve healthcare and enable the study of personalized medicine and treatment information. In particular, three prominent use cases for ANN in healthcare are medical imaging, digital pathology, and omics that are often employed together to increase the accuracy of diagnosis.

In this work, we studied AI-based techniques, especially focusing on Deep Learning (DL) approaches, to support classification, analysis, and diagnosis in the biomedical domain and to more transparent and explainable result process achievements. After a general introduction, AI-based approaches are described, along with their application in several biomedical domains (e.g,. vessel segmentation, clinical data classification, etc.); furthermore, a study of the DL decision-making process with the aim of providing model explanations is presented.

More in detail, the remainder of this manuscript is structured as follows.

- In the first part, we provide a general introduction about the main concepts of this thesis. Such part is divided into three chapters; in the first, we present a detailed description of electronic health record and omics data, respectively, by highlighting the main characteristics and issues and their effectiveness in improving disease prediction and prevention. We describe some existing and emerging applications via clinical and omic data; the second chapter describes the principle concepts of DL: we provide a definition and description of the various ANN-based techniques, along with different examples of their application via medical imaging; in the last chapter, we discuss the scope of this thesis.

- The second part is devoted to the proposal of two DL-based approaches applied in biomedical analysis. In the first chapter, DL techniques will be applied to segment the vascular tree in ilio-femoral district. In the second chapter, we present the application of an object detection and segmentation network, namely Mask R-CNN, to perform instrument instance segmentation in laparoscopic videos. The approach relies on the use of raw images, binary segmentation and optical flow to improve the multi-instance segmentation task.

- The third part reports on a proposal for the combination of data visualization and data reduction techniques, for the manipulation and the conversion of clinical and omics data into images to perform DL-based classification.

- Eventually, in the fourth part we perform an accurate analysis of the internal processes performed by the networks during the classification tasks with the aim to highlight the most important elements involved in the training process and to validate the network's decisions and clinical suggestions.

PART I

Context and State of the Art

CHAPTER 1

Clinical and Medical Data

Contents

1 Clinical data

The expression "health-related information" usually refers to clinical data, and it is associated with regular patient care or it comes as a part of a clinical trial program. Clinical data represent an important resource to enable and guide the acquisition of novel knowledge (e.g., to provide outcome and therapies predictions) and best practices (e.g., to ensure data completeness, reliability, and correctness) in healthcare [McGinnis et al. 2011]. Clinical data are also important to find timely treatments and appropriate care to the patient, and to enable a (sort of) self-learning system to continuously improve quality of care. Clinical data are collected and translated into Electronic Health Records (EHRs); they include administrative and demographic information, diagnoses, treatments, prescription drugs, physiologic monitoring data and other health information, as reported in Section 1.2. The use of EHRs in clinical research studies provides several advantages, such as defining the most suitable treatments, reducing hospitalization cost, and providing personalized medicine. EHRs are used for various data science and ML studies, including risk prediction for breast cancers [R. Li et al. 2020], statistical analysis of diseases, diagnosis of rare pathologies [Garcelon et al. 2020] and classification of heart disease and diabetes [Brisimi et al. 2018].

In the last decade, several studies have shown that a proper analysis and combination of omics and EHRs data can help in discovering relevant information and distinctive attributes related to specific diseases [Wise et al. 2019; Bruno et al. 2019; B. Zhu et al. 2017; Oromendia et al. 2020]. Explicitly, omics studies are used to analyze types of molecules in samples which can be measured in terms of character and quantity, with the aim of investigating the patterns or relations to the sample attributes [Yamada et al. 2020]. The comprehensiveness of these data allows to generate or confirm hypotheses on biological or medical conditions [X.-T. Yu et al. 2018] and to provide a basis for the precision medicine [P.-Y. Wu et al. 2016].

The precision health, an evolution of the precision medicine, combines omics data with lifestyle, clinical data, and environmental factors. These data can be used to identify and predict disease diagnosis, treatment, prevention and individualized early diagnosis [Fu et al. 2020]. Precisely, the personalization of medical treatment based on specific patient characteristics is possible by improving the understanding of the physiological and biological mechanisms of disease, responsible for the

spreading of omics data, and by developing patient-based algorithms [Madhavan et al. 2018]. Thanks to these advances, the attention on precision health is growing quickly, thereby allowing an improvement of outcomes and reducing unnecessary treatment. However, clinical data as well as omics data present several issues that can afflict study results, especially in terms of quality, such as incompleteness (e.g., missing information), inconsistency (e.g., information mismatch between various or within the same data source) and, inaccuracy (e.g., non-specific, non-standards-based, inexact, incorrect, or imprecise information) [Ford et al. 2020].

Omics Data Management (ODM) and Clinical Data Management (CDM) play a crucial role in generating high-quality, reliable, and statistically significative data from clinical trials. ODM is used to address the uncertain or unexpected findings, that may have relevant impacts in medical diagnosis (e.g., the collection and processing of genome data that are not sufficiently standardized or valid). CDM provides high-quality data by reducing the number of errors and missing data to improve data analysis [Krishnankutty et al. 2012].

Generally, data management is the process of collecting, cleaning, and processing data according to standard rules. This process requires specific software applications and best practices to ensure data completeness and reliability. Section 1.1 and 1.2 report a detailed description of omics data and electronic health record, respectively, by highlighting the main characteristics and issues.

Part of the work proposed in this chapter has been accepted and it will be published in the book project "Artificial Intelligence in Medicine", edited by Niklas Lidströmer and Hutan Ashrafian and published by Springer.

1.1 Omics data

Understanding of human health and diseases requires a proper interpretation of molecular interactions and variations at multiple levels such as genome, epigenome, transcriptome, proteome, and metabolome [Subramanian et al. 2020], which together go under the name of omics data. Specifically, omics refers to the collective technologies used to explore roles, relationships, and actions of the different types of molecules composing organism cells [Gajula 2012] and, potentially, responsible for specific condition or disease [Tebani et al. 2016]. The omics data can be classified in different types according to the type of experimental data.

- **Genomic** is a technique used to improve diagnosis through identification of genomic conditions, to improve clinical management, prevent complications, and promote health [Wise et al. 2019]. The genome is the complete sequence of DNA in a cell or organism.

- **Transcriptomic** is used to measure the chemical states of DNA and its binding proteins, RNA, and metabolites, respectively [Yamada et al. 2020]. The transcriptome is the complete set of RNA transcripts from DNA in a cell or tissue. The information of an organism is stored in the DNA of its genome and expressed through transcription, that is also used in disease diagnosis and profiling [Lowe et al. 2017].

- **Epigenomic** studies the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. The epigenome consists of reversible chemical modifications to the DNA. Potentially epigenetic patterns can be useful biomarkers to detect cancer cells and to classify different disease types [Weichenhan et al. 2020].

- **Proteomic** is used to study proteins, vital parts of living organisms with many functions. The proteome is the complete set of proteins expressed by an organism, tissue or cell. Proteomic tecniques enable the detection and quantitation of protein profiles associated with the disease state [Clark et al. 2020].

- **Metabolomic** monitors and studies the presence and the concentration of specific metabolites associated with a particular disease. The metabolome is the complete set of small molecule metabolites found within a biological sample. This methodology helps to identify clinically relevant biomarkers as it best mirrors the human phenotype [Njoku et al. 2020].

Using high-throughput data acquisition such as Next Generation Sequencing (NGS) and Mass Spectrometry (MS) allows to perform fast accumulation of omics data. These data, after a cleaning process, can be used in determining medical biomarkers, extracting molecular profiles, identifying statistically significant molecules, or defining models able to explain molecular interactions in a specific context. Specifically, the genomics approaches have been used to identify the genes and genetic loci

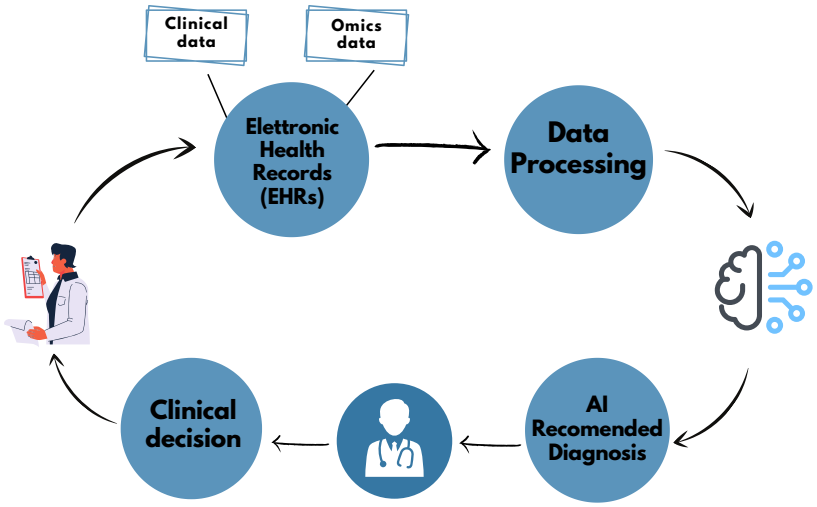


FIGURE 1.1: Workflow of data processing steps. Data cleaning and dimensionality reduction are performed on raw data. Sequence mapping and normalization are usually used to handle the complexity of omics data.

involved in the development of human diseases [Iacobucci et al. 2019], the epigenomics to find epigenetic markers [Soler-Botija et al. 2019], the proteomics for proteins and peptides [Taha et al. 2019], and the metabolomics for low-abundance metabolites [Shao et al. 2019]. These data are composed of a huge amount of data points, requiring standard data formats and publication guidelines [Chervitz et al. 2011] to improve data sharing, acquisition, analysis, and usage. The standard formats ensure unambiguous communications, clear experimental designs, treatments, and analyses to support the conclusions, guarantee an independent reproduction [Chervitz et al. 2011] and facilitate the creation of standardized public data repositories. The current state-of-the-art counts several methodologies to facilitate a good exchange and integration of data. The Minimum Information About a Microarray Experiment [Brazma et al. 2001] is used to define a guideline for the minimum information required to describe a DNA microarray-based experiment, the Minimum Information About a Proteomics Experiment [C. F. Taylor et al. 2007] to define guidelines for proteomics studies, and the Minimum Information about a high throughput SEQuencing Experiment [Kahl 2015] to define guidelines for sequencing study.

1.2 Electronic Health Records

EHRs construct the digital version of a person's medical information and history, preserved over time by a healthcare provider. They are used to timely and consistently collect patient information providing more extensive and accurate clinical care. EHRs are also used to predict future outcomes of specific patient based on both individual-related and population-related data. In fact, a proper use of EHRs can improve healthcare quality, including benefits such as a secure long-term storage, consistency, standardization, and accessibility of patient information [Howe et al. 2018].

EHRs can be divided in **structured** and **unstructured** data. Demographics, medical and surgical histories, diagnoses and procedures, patient examinations, and results from various clinical studies, are example of structured data. While clinical notes (e.g., hospital admission notes, physical therapy and history and physical examination) belongs to unstructured category.

The structured data use a uniform format in the EHRs system with a controlled vocabulary and predetermined values that make these data consistent and easily extractable. On the other hand, the unstructured data do not use a standard format: healthcare providers can include free text such as additional health information and details on clinical examinations. It means that the same clinical information can be recorded in EHRs in different ways depending on the user, making the analysis difficult because of the presence of (1) heterogeneous data formats, (2) abundant typing and spelling errors, (3) violation of natural language grammar, and (4) rich domain-specific abbreviations, acronyms, and idiosyncrasies [P.-Y. Wu et al. 2016]. The unstructured data requires the use of additional tools, such as natural language processing (NLP), to standardize, codify, and extract relevant information [Abul-Husn et al. 2019].

1.3 Diagnostic-related information and treatment information

Diagnostic-related information can benefit patients in receiving the right treatment, helping healthcare operators to provide the most appropriate preventive interventions and therapeutic strategies [Wurcel et al. 2019], avoiding or shortening hospitalization, reducing inappropriate use of drugs and improving using resources [Abul-Husn et al. 2019]. Diagnostic information can help clinicians to identify certain

biomarkers in the body. However, this approach does not allow any insight into the interaction among the disease information and the environment responsible.

EHRs combined with the emergence of omics data offer novel research perspectives and opportunities in health systems to improve the health management and enable the study of personalized medicine and treatment information. In this scenario, using omics data becomes relevant helping therapy customizing according to the patient molecular profiles. These data are a crucial element predicting biomarkers, determining predisposition, diagnostic and prognostic, and identifying risk factors to improve the traditional symptom-driven practice of medicine [Ahmed 2020].

Extracting knowledge from large and complex clinical and omics data is a challenging task. ML algorithms can overcome these issues thanks to their ability to adapt to specific settings and omic types [Zampieri et al. 2019]. During the last years, several researches have been proposed to examine the ML and DL application in the diagnostic-related and treatment information area. Some existing and emerging applications are described in Section 3.

2 Data Processing

Omics and EHRs data can be difficult to analyze and computationally expensive to process due to their high-dimensionality. In particular, omics data are composed of many dimensions/features much larger than the number of samples available, while EHR data usually contains a large sample size of high-dimensional data, but each individual sample is sparsely populated [P.-Y. Wu et al. 2016]. Information quality problems could make these data not really suitable for clinical research. Data quality assessment and dimensionality reduction are necessary to remove irrelevant and redundant data, and to preserve the characteristics of the original data. An overview of the data processing operations is shown in Figure 1.1.

In the present section, we first describe techniques used to handle missing values in Section 2.1. Then, we illustrate dimensionality reduction technique in Section 2.2. Eventually, we present further pre-processing approaches that are used to deal with omics data in Section 2.3.

2.1 Missing value imputation

The quality of each omic and EHR data must be carefully assessed to ensure measurements reproducibility and to maximize the data analysis performance. Over the last decades, different approaches were proposed to tackle this issue. One of the most common approaches is called "single imputation", which relies on the removal of the missing rows, by ignoring subjects with incomplete information or replacing the missing items with plausible values (e.g., means of the observed cases). The single imputation approach presents some limitations caused by the high proportion of subjects discarded [Voillet et al. 2016], the distortion of the distribution of the variables, and the introduction of additional biases [Jakobsen et al. 2017]. Moreover some state-of-the-art approaches such as multiple imputation [Jakobsen et al. 2017], inverse probability weighting [L. Liu et al. 2018], expectation-maximization [Malan et al. 2020], multivariate imputation by chained equation [Zhongheng Zhang 2016] were considered more robust than single imputation.

2.2 Dimensionality reduction

Feature extraction and selection methods are used to maximize performance analysis and improve result understandability. **Feature Selection** is used to select features in input that contains relevant information for solving a specific problem. **Feature Extraction** transforms the input space into a low-dimensional subspace that preserves the most relevant information [Khalid et al. 2014].

Feature selection techniques can be part in filter, wrapper, or embedded methods. In the **filter** approach each feature is evaluated individually using its general statistical properties [Lazar et al. 2012], making the approach faster, without explicit class labeling [P.-Y. Wu et al. 2016]. The most common filter methods are set out below. "Mutual Information" measures the level of dependence between two random features (i.e., the amount of information that variable V1 knows about another variable V2) [Vergara et al. 2014]. "Information Gain" evaluates the gain of each feature to a certain class, by computing the entropy (i.e., level of impurity), such that the feature with higher information is the much related, while the unrelated feature offers no information [Hira et al. 2015]. "Minimum Redundancy Maximum Relevance" [Lazar et al. 2012] iteratively selects features with the maximum relevance,

decreasing the redundancies within each class, such that, any mutually exclusive features are selected [Peng et al. 2005].

The **wrapper** approach uses learning techniques to select the optimal feature subset. It typically requires high computational costs and is affected by an overfitting risk, but it shows better performances than the filter approach [Almugren et al. 2019]. Genetic Algorithm, Ant Colony Optimization and Particle Swarm Optimization are examples of wrapper approaches inspired by natural evolution, ant colonies and flocks of birds, respectively, to generate a population of features that optimize the solution.

The **embedded** methods integrate ML algorithms with recursive feature elimination [P.-Y. Wu et al. 2016]. These methods are less computationally expensive and less prone to overfitting than the wrapper ones. An example of an embedded method is Support Vector Machine (SVM), a supervised ML algorithm used to search for a hyperplane that optimally divides the tuples from one class to another. Therefore, SVM allows us to identify the main features used in classification and remove the not important ones [Pal et al. 2010].

Among the various feature extraction techniques, **Principal Component Analysis (PCA)** is a commonly used method. PCA is an orthogonal linear transformation that converts variables into a new smaller set, called principal components. The number of principal components is less than or equal to the number of original feature variables [L. Yang et al. 2019].

Considering a multivariate data matrix X, with n rows and j columns, containing the measurements of sample n on variable j (e.g., a matrix composed of n patients and j genes), PCA summarizes the information in X by defining a new matrix X_1 of dimensions N_1 x J computed as follows [Hoefsloot et al. 2020]:

$$X_1 = T_1 + P_1^T + E_1 \tag{1.1}$$

where N_1 is the total number of samples in X_1, J is the total number of variables in matrix X_1, T_1 represents the matrix N_1 x R_1 containing the scores of the model of X_1, P_1 is the matrix J x R_1 containing the loadings, E_1 is the matrix N_1 x J with the residuals of the model and R_1 is the number of components selected for the PCA model of X_1. The final goal of PCA minimizes the sum of squares of E_1. The size of

the resulting set is smaller than the size of the original.

2.3 Different omics processing

Normalization (I) and sequence mapping (II) could be other appropriate pre-processing approaches to handle the complexity of omics data.
(I) **Normalization** is important in removeìing unwanted systematic bias while maintaining real biological differences in the observed datasets. In this context, different methods based on several statistical models (e.g., unit norm, median, and quantile), scaling methods (e.g., auto-scaling, range scaling, Pareto scaling, vast scaling, and level scaling), and data transformation (e.g., log and power) [Sugimoto et al. 2012] have been used. Other approaches like Locally Weighted Scatterplot Smoothing algorithm [Cleveland et al. 1988] are used to normalize the samples by adjusting their variability. It removes the bias in the data which frequently deviate from zero for low-intensity spots. Normalization also enables the comparison of different samples in MS. Normalizing spectrum makes possible to identify and remove sources of systematic variation in MS data depending, for example, on variation in samples amount or in the detector sensitivity.

(II) **Sequence mapping** is the process of comparing millions of sequences generated, for instance, by NGS against the reference genome to obtain one alignment between each read and the genome. In NGS and MS analysis, mapping is fundamental. It is the basis for further analysis, for instance, to estimate the abundance of transcripts and variant detection [Thankaswamy-Kosalai et al. 2017]. The common NGS mapping tools are based on a hash table or index-based algorithms, while heuristic-based aligners are demonstrated to be less expensive in terms of computation, such as Genomic Mapping and Alignment Program [T. D. Wu et al. 2005] and Burrows-Wheeler Alignment [H. Li et al. 2009].

3 Existing and emerging applications in Medical Diagnosis

Recent improvements in ML and DL approaches provided useful techniques to recognize patterns from clinical and omics data and to predict patients future outcomes. A conceptual model to illustrate the diagnostic process is shown in Figure 1.2. Several works have already proved the validity of these methods in providing improved

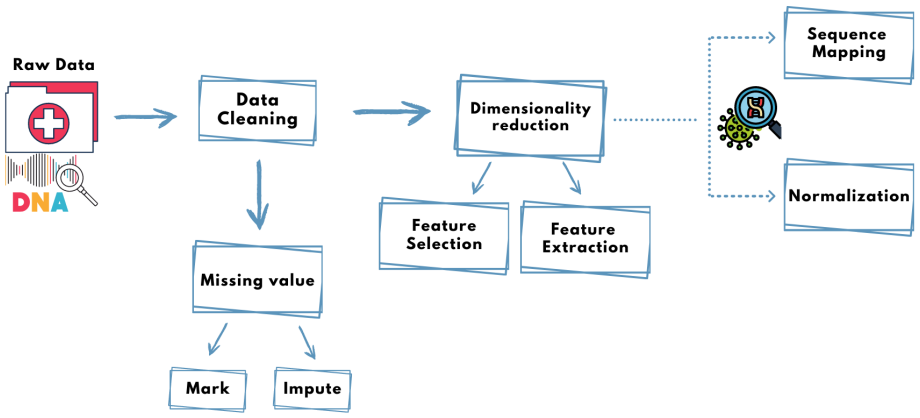


FIGURE 1.2: EHRs are composed of clinical data and omics data that are collected during medical examinations. Data processing operations are performed to prepare data for AI techniques. The output is used by healthcare providers to improve the diagnosis and provide a clinical decision.

and more generalizable risk prediction models. These algorithms make use of high-dimensionality data, such as EHRs or omics data, to determine the combinations of variables able to predict a reliable outcome.

In this section, we will describe the latest advances in ML and DL on omics data and EHRs. Specifically, omics data are considered from the molecular profiles perspectives: genomics, transcriptomics, epigenomics, proteomics and metabolomics.

3.1 Omics data application

Genomic data. DL technologies can be used to predict gene expression and to investigate interaction between genomes and diseases. Chen et al. [Y. Chen et al. 2016] made use of a DL method to infer the expression of target genes from the expression of landmark genes. They used the microarray-based Gene Expression Omnibus dataset to train the model, achieving a performance significantly better than logistic regression, with an improvement of 15.33% on mean absolute error. Chen et al. [R. Chen et al. 2020] presented a DL-based approach to explore genomes and disease relationships. The authors performed jointly supervised classification, unsupervised clustering, and dimensionality reduction on high-dimensional genomic data to identify cancer subtypes. This approach was tested on breast cancer and outperformed the existing methods, identifying more robust subtypes using fewer genes. A limitation of the work is represented by the generalization capability. The performance decreases in predicting other types of cancers where molecular subtypes have not

yet been well established.

Transcriptomic data. DL technologies can be also used to identify the association between RNA and disease. Thomas et al. [Thomas et al. 2017] proposed a DL method to perform the classification of pre-miRNAs. This approach is composed of an unsupervised stage with hidden layers pre-trained as restricted Boltzmann machines and it is followed by a supervised tuning of the network. The approach achieved a level of accuracy of 0.97. Bobak et al. [Bobak et al. 2019] presented a data analysis framework that directly integrates multiple publicly-available expression array datasets in order to identify a more reliable gene signature for the diagnosis of tuberculosis. The authors evaluated different ML algorithms including random forest, support vector machine with the polynomial kernel and partial least square discriminant analysis. According to the analyses, the authors proved that the best result was obtained using random forest with an accuracy value of 0.95.

Epigenomic data. DL methods were also used in the context of predict epigenetic effects of DNA sequence alterations (such as chromatin accessibility, DNA methylation and histone modifications). Quang et al. [Quang et al. 2016] proposed a hybrid approach based on Convolutional Neural Network (CNN) and Long Short-term Memory (LSTM) to predict the chromatin effects of non-coding DNA sequence alterations. This approach simultaneously learn motifs and a complex regulatory grammar between the motifs. The authors showed that the approach outperforms other methods for predicting the properties and function of DNA sequences across several metrics, including Area Under the Curve (AUC). A limitation of the Quang et al. approach is the dimension of sequence data used to train the network; indeed, it can only process sequences of constant length with static output. Yin et al. [Yin et al. 2019] proposed a DL approach for integrating sequence information and chromatin data to perform prediction of modification sites specific to different histone markers. The architecture is composed of three modules corresponding to a DNA sequence, chromosome accessibility and a joint module, respectively. The approach outperformed several baseline methods in a series of comprehensive several validation experiments. The authors did not incorporate recurrent neural network architecture, such as long short-term memory units that may improve the performance by exploiting the sequential natural DNA fragments.

Proteomic data. DL technologies are also effective in the identification of protein structures and protein contact map prediction. Wang et al. [Sheng Wang et al. 2017] presented a DL method to improve protein contact prediction by integrating both evolutionary coupling and an ultra-deep neural network to preserve sequence information. The approach achieved the highest F1 score on protein structure prediction. Liang et al. [Liang et al. 2019] proposed DL algorithms with stacked autoencoders for the analysis of FLT3-ITD mutation in acute leukemia patients. The authors made use of dimensionality reduction algorithm to reduce the number of proteins and their approach achieved an accuracy of 0.97.

Metabolomic data. Finally, DL technologies can be used to capture the metabolic features of complex traits and predict metabolic pathways. Stamate et al. [Stamate et al. 2019] relied on several state-of-the-art algorithms, such as DL, Extreme Gradient Boosting and Random Forest to predict Alzheimer's Disease versus cognitively normal with metabolites as predictors. The implemented framework captured metabolic complexity in Alzheimer's disease, achieving an AUC value of 0.88 and 0.85 using the Extreme Gradient Boosting and DL-based approach, respectively. A limitation of the study is represented by the small size of the cohort. Baranwal et al. [Muzio et al. 2020] proposed to use of the Random Forest classifier for metabolic pathway prediction. The input of the classifier was extracted from molecular structures in SMILES format using Graph Convolutional Networks. This approach achieved 95% accuracy in predicting metabolic pathways, showing a great ability in estimating the relative contribution of metabolites in distinguishing pathway classes.

3.2 Clinical data application

Several works have been proposed to perform diagnosis prediction using clinical data. Corey et al. [Corey et al. 2018] defined different ML algorithms including penalized logistic regression, random forest models, and extreme gradient boosted decision trees to identify high-risk surgical patients from EHRs. The experimental analysis showed that the best result was produced by penalized logistic regression models with an AUC value of 0.92. Miotto et al. [Miotto et al. 2016] proposed a DL-based approach composed of a stack of denoising autoencoders to process EHRs in

an unsupervised manner to capture stable structures and regular patterns in a variety of clinical risk-prediction tasks (e.g., diabetes mellitus with complications, cancer of rectum and anus). The proposed approach outperformed state-of-the-art methods based on raw EHR data (i.e., no feature learning applied to EHR data). A limitation of these approaches is that the authors did not include on temporal information to improve the performance of predictive models, making them not applicable in real clinical settings.

To tackle time dependencies in EHRs, Rohit et al. [Kate et al. 2019] presented a continual prediction framework based on logistic regression to predict Acute kidney injury (AKI) before the development of disease at any time during the hospitalization. In particular, the approach continually predicts over the entire hospital stay whenever any patient variable changes (e.g. AKI occurrence). A limitation of the work is related to the data used to predict the disease: the authors used only the structured part of EHR, discarding the unstructured component. Mehak et al. [Gupta et al. 2019] proposed a Recurrent Neural Network architecture with LSTM, which learns the patient representation from the temporal data collected over various visits of the patient, to predict future obesity patterns from children's medical history. Similarly, Jeong et al. [Lee et al. 2020] proposed a Multi-scale Temporal Memory to model clinical events at different time scales in EHRs. Thanks to this approach, information about past events on different time-scales are collected and used to perform on-the-fly prediction. The authors showed that the proposed method, combined with different patient states, can cover different temporal aspects of patient states, achieving an accuracy improvement of 4.6% than baseline approaches and of 16% the prediction based on LSTM approach. A limitation of these approaches is related to the limited size of the observation windows in the EHRs data. Finally, Che. et al. [Che et al. 2018] developed a variation of the recurrent GRU cell (GRU-D) able to handle missing values in clinical time series by incorporating time intervals inside the architecture. The authors showed an improvement of AUC on two real-world health care datasets for classification and mortality prediction tasks.

CHAPTER 2

Deep Learning

Contents

1 Introduction

The advent of high-dimensional data has required ML to solve complex biological problems in the real world and to enable the automation of processes that would have been impossible to do through traditional computer science techniques and algorithms [Roy 2015]. ML is a subset of artificial intelligence, which enables systems to automatically learn and improve from experience without being explicitly programmed. Mitchell [Mitchell et al. 1997] defines ML as follows: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". For instance, in the domain of medical image classification, the task T is to classify medical images into different categories, the performance measure P is the percent of the images correctly classified, and the experience E is a dataset containing medical images used to train and test the algorithm [J. Wang 2003]. These algorithms aim to identify (to learn) a function f to map the input data X into output prediction Y. The function f depends on the type of the learning algorithm chosen. According to the nature of the training data, four main categories of algorithm exist: (i) supervised learning, (ii) unsupervised learning, (iii) semi-supervised learning and (iv) reinforcement learning.

- Supervised learning techniques construct predictive models by learning from a large number of training examples. Each training example has a label indicating its ground-truth output, then the result is known. In a more formal way, we can define the training set as $\Delta =< x, y >$ in which each element $< x, y >$ is an input-output pair. The input variables x are also called features while the output variables y labels or targets. Δ is used to find a deterministic function able to map any input to an output. According to the expected output variables, supervised learning tasks can be subdivided into **Classification** and **Regression** [Lizhi Wang et al. 2021].

    1. **Classification** is a predictive model used to approximate a mapping function from input variables x to discrete output variables; these variables can be labels or categories. Depending on the number of classes that we want to predict, classification problem can be binary (2 output classes) or multi-class ($>$ 2 output classes).

2. **Regression** is used to predict a numerical value given some input variables x (e.g., predicting air temperature on a specific day or future prices of securities) [Bengio et al. 2017].

- Unsupervised learning techniques create the model by deducing structures in the input data not (previously) class labelled. Then, the result is unknown. The training set is expressed in the form of $\Delta =< x >$, where x represents the input variables and no label variables are included. For instance, **clustering** is the process of partitioning a set of data points according to some measure of similarity (e.g., distance) and is one of the most widely used unsupervised learning techniques [Alashwal et al. 2019].

- Semi-supervised learning techniques make use of the combination of the unlabelled data available and the labeled data to train a model [Bengio et al. 2017].

- Reinforcement learning refers to ways of improving performance through trial-and-error experience [Barto 1997]. The reinforcement learning algorithms interact with an environment, defining a system of feedback loop between the learning system and its experiences [Bengio et al. 2017]. Specifically, an agent interacts with the environment by performing actions and it receives a reward that indicates how good the action is [Kress 2010]. The optimal performance is defined by selecting the actions that maximize the reward.

In the last decades, several methods based on ML techniques were developed to perform various clinical tasks, such as automatic identification of pathological condition or classification of patient diagnosis. In this context, DL methods have obtained outstanding results in the areas of image processing and computer vision [Xizhao Wang et al. 2020], paving the way for several breakthroughs and improvements in many real-life contexts, such as automatic personalization of medical treatment.

In the present chapter, we provide an overview of the main techniques used in this research work. We describe DL and the differences between the various existing algorithms, including object detection networks and their applications in biomedical domain.

2 Artificial Neural Networks

In the last decade, ANNs, which are ML-based approaches, are widely used to solve many real-world problems. ANNs are computational networks inspired to human brain that are usually modeled using neurons. ANNs are modeled using layers of artificial neurons, or computational units, where each unit takes a number of real-valued inputs and produces a single real-valued output [Mitchell et al. 1997]. In details, neurons are organized in layers, i.e., the input layer, which represents a specific input data vector, the output layer, which produces the final result and one or more hidden layers, used to connect the input and output layers. Specifically, the hidden layers perform nonlinear transformations of the inputs entered into the network and produce an output through an activation function. The number of hidden layers can vary according to the task the network has to deal. The simplest ANN is the perceptron, developed in the 1958 by the scientist Frank Rosenblatt [Rosenblatt 1958]. A perceptron takes several binary inputs (i.e., $x_1, x_2, \ldots, x_n$) and produces a single binary output. In order to compute the output, a perceptron uses real numbers, called weights (i.e., $w_1, w_2, \ldots, w_n$), to express the importance of the respective inputs to the output. The neuron's binary output (i.e., 0 or 1) is determined as follows:

$$output = \begin{cases} 0, & \text{if } \sum_{n=1}^{N} (x_i, w_i) + b \leq 0 \\ \\ 1, & \text{if } \sum_{n=1}^{N} (x_i, w_i) + b > 0 \end{cases} \tag{2.1}$$

where b is the bias term.

The weights and biases are adjusted according to the perceptron learning rule with the aim to minimize the difference between the actual and the desired outputs [Wallisch et al. 2009]:

- If the output is correct, the weight vector associated with the neuron is not changed.
- If the output is 0 and the expected value is 1, the input vector is added to the weight vector.
- If the output is 1 and the expected value is 0, the input vector is subtracted from the weight vector.

The weight vector is changed to point more toward the targets; indeed, the correction performed on the weights is proportional to the distance between predictions and targets. The procedure is repeated until the algorithm converges to a possible solution, finding a model able to correctly classify the largest possible number of the examples. Several optimization techniques could be applied to learn the weights of the model. In the following section, we will provide an overview of gradient descent: one of the most used optimization approaches.

2.1 Optimization

The gradient descent is one of the most popular algorithms to perform optimization in supervised learning [Bengio et al. 2017]. It is an iterative optimization algorithm used to find the minimum value of a function by moving in the direction of steepest descent as defined by the negative of the gradient. We can think of the search space as a valley and, starting at the top of the mountain, we can first move downhill in the direction specified by the negative gradient. Next, the negative gradient is computed on the new point which is moved again to the specified direction. This process is repeated until the local minimum or the bottom of the graph is reached.

The learning rate modulates the size of these steps. The actual size is generally proportional to learning rate and it is adjusted by the optimization algorithms. Specifically, a high learning rate may help preventing local minima, as the solution may jump out of the minimum. Higher learning rate can be used for large batch size while small learning rate can be used for small batch size might to maintain stability due to the high variance in the estimate of the gradient [Bengio et al. 2017]. Generally, learning rate is decreased during the training to avoid oscillation and getting stuck in undesirable local minima. If the learning rate is too high, the training may not converge.

Define the proper starting point (i.e., the randomly values used to initialize the weights) and the learning rate is not an easy task, making the training difficult. Several variations to the gradient descent were proposed.

1. **Stochastic Gradient Descent (SGD)** [Tieleman et al. 2012] [Ketkar 2017] performs the update of the model parameters for a batch of training examples, reducing the training time. Also, the stationary points could be different in each epoch, allowing the algorithm to escape from local minima.

2. **Root Mean Square Prop (RMSprop)** adapts the learning rate for each parameter as SGD but it uses the squared gradients to scale the learning rate.

3. **Adam** [Kingma et al. 2014], similarly to RMSprop, computes individual learning rates for different parameters. It estimates the first and second moments of gradient to adapt the learning rate for each weight of the neural network.

2.2 Backpropagation

Backpropagation was proposed to generalize perceptron learning algorithm to the multi-layer network where the loss is defined as a composition function of the weights in earlier layers. The gradient of this cost function can be computed using backpropagation algorithm [Bengio et al. 2017].

This algorithm tries to find the minimum of the error function in the weight space using the method of gradient descent. This error helps in determining how far the network is from the correct prediction on the training set [Bergel 2020]. Specifically, backpropagation advantages the chain rule of differential calculus, which computes the error gradients in terms of summations of local-gradient products over the various paths from a node to the output [Aggarwal et al. 2018]. The heart of backpropagation is the partial derivative computed on cost function C w.r.t. any weights and biases in the network. Partial derivatives allow to update the parameters (i.e., weight and bias b) with the aim of minimizing C [Nielsen 2015].

Backpropagation is composed of three main steps [Bergel 2020]:

1. **Forward Propagation**: The information is propagated in the forward direction through the network to compute values from data input to output for each neuron.

2. **Backward Propagation**: The output produced in the previous step is compared to the actual training dataset, computing the error. This error is propagated backforward from the right-most layer (i.e., the output layer) to the left-most layer (i.e., the first hidden layer) according to the chain rule [Bergel 2020].

3. **Update weight values**: The weight and bias of each neuron are adjusted to reduce the overall error made by the network, according to the gradient of the error computed in the previous step.

2.3 Activation Functions

The activation functions are an important component in DL, providing considerable effects on the ability of neural networks to converge and accelerating convergence speed. Activation functions perform a nonlinear transformation on the input to achieve better results on a complex neural network. They basically decide which neurons will be activated or deactivated to obtain the expected output. Several activation functions were implemented.

1. **Sigmoid**:

$$\sigma(x) = \frac{1}{(1 + e^{-x})} \tag{2.2}$$

   The sigmoid non-linearity takes a real-valued number and projects it into range $[0, 1]$ [Nielsen 2015]. This function saturates when the argument is very positive or very negative. This behaviour implies that sigmoid becomes very flat and insensitive to small changes in the input [Bengio et al. 2017].

2. **Hyperbolic tangent (Tanh)**:

$$Tanh(x) = \frac{1 - e^{-2x}}{(1 + e^{-2x})} \tag{2.3}$$

   The Tanh non-linearity takes a real-valued number and projects it into range $[-1, 1]$ [Nielsen 2015].

3. **Rectified Linear Unit (ReLU)**:

$$ReLU(x) = max(x, 0) \tag{2.4}$$

   ReLU is a non-linear function and its capability of backpropagating the errors and accelerating the convergence of stochastic gradient descent w.r.t. the sigmoid and Tanh functions makes it one of the most popular activation functions [J. He et al. 2018]. However, since the negative values are converted into zero, the weights in that region are not updated during backpropagation, potentially resulting in dead neurons that never get activated.

4. **Softmax**:

$$sofmax(x) = \frac{e^{x_j}}{\sum_i (1 + e^{x_i})} \tag{2.5}$$

The softmax function projects the outputs for each class into range $[0, 1]$ and the sum of the outputs is always 1. Softmax is usually used at the last layer (i.e output layer) to classify more than two classes; indeed, the previous activation functions produce a single output for a single input while softmax generates multiple outputs for an input array.

3 Deep Learning Architectures

In recent years, DL technologies gained a lot of popularity, due to their impressive results in the fields of image processing, pattern and object recognition [H. Wu et al. 2019]. DL has been mainly employed in several research areas. Among these, the ones of interest here are the following:

- **Image classification** refers to the task of assigning a label to a given image or outputting a probability that the input belongs to a particular class.
- **Image segmentation** is the process of partitioning an image into multiple segments based on the pixels characteristics to locate objects and boundaries in images [Narayanan et al. 2015].
- **Semantic segmentation**, similarly to image segmentation, refers to the process of linking each pixel in the given image to a particular class label. However, it treats multiple objects of the same class as a single entity.
- **Object detection** refers to the identification and correctly labeling all the objects present in the input image. This approach relies on (i) *object localization* to locate exactly the position of the object in the image through bounding box or enclosing region and (ii) *image classification* described above.
- **Instance segmentation** requires the prediction of object instances and their per-pixel segmentation mask, to assign different labels for separate instances of objects belonging to the same class and provide their corresponding segmentation [Hafiz et al. 2020].

In this section, we will present a general overview of the main DL approaches, focusing on a detailed description of the methods used to address the specific problems presented in this manuscript.

3.1 Convolutional Neural Networks

CNNs represent a huge breakthrough in image recognition, due to its capability in exploiting spatial or temporal correlation in data. The structure of CNNs is made of multiple learning stages composed by a combination of convolutional layers, non-linear processing units, and subsampling layers [Khan et al. 2019]. CNN relies on backpropagation algorithm to update weights according to the target objective and a multilayered, hierarchical structure which enables the extraction of low, mid, and high-level features. Indeed, CNN is able to process the information at different scales, with small features extraction first. This is similar to the functioning of the human visual system.

CNNs are architecturally invariant to translation thanks to the convolution and/or pooling operations they are endowed with [Biscione et al. 2020]. It means that the network is able to recognize patterns in every possible position (i.e., shifted, tilted or slightly warped within images). In detail, this ability is achieved thanks to three main properties of CNN architecture:

1. **Local Receptive Fields**: in CNN each neuron considers only a specific region of the input data called local receptive field which determines the association from an output feature to an input region. Local receptive fields are used to slide across the entire input image by a certain quantity of pixels, allowing neurons to learn patterns such as lines, edges and small details of image.

2. **Shared Weights**: In CNN the weights and bias values are the same for all hidden neurons in a given layer. Then, shared weights allow the detection of exactly the same feature at different locations in the input image (i.e., translation invariance principle) [Nielsen 2015; Hawkes 2004]. The mapping from the input layer to the hidden layer is called Feature map (or filters). CNN architectures are composed of several feature maps necessary to perform image recognition [Nielsen 2015].

3. **Pooling layer** also called downsampling, aims to reduce the dimensionality of the feature maps to simplify the information from the convolutional layer. This process is done is according to a specific factor called stride.

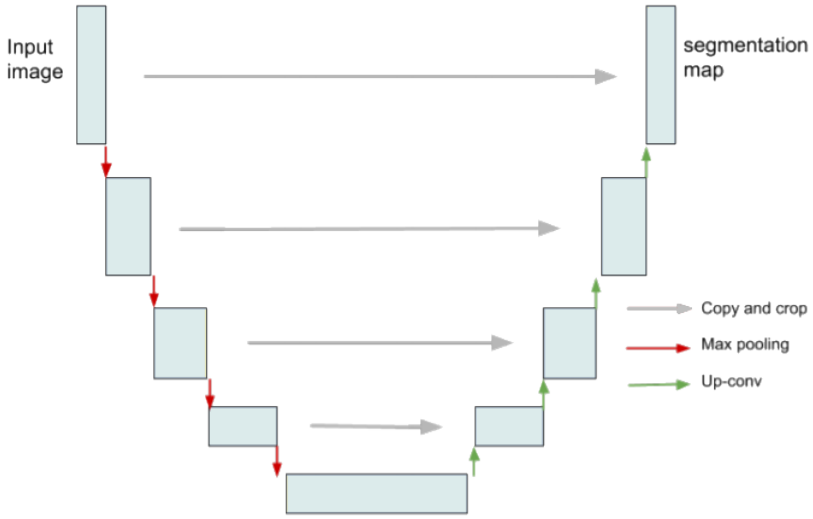


FIGURE 2.1: U-net architecture inspired by [Ronneberger et al. 2015], with the contracting path on the left and the expansive path on the right. The max pooling operations decrease the sizes of the feature maps while increase the number of feature channels. In the expansive path an upsampling followed by convolution layer halves the number of feature channels, keeping the same dimensionality as the input for the final output. Grey arrows indicate long skip connections between layers in encoder and decoder which is used to recover fine-grained details in the prediction [Ronneberger et al. 2015].

3.2 U-net

U-net is a CNN architecture for fast and precise segmentation of images [Ronneberger et al. 2015]. It was first designed for medical image segmentation [X. Li et al. 2018; Z. Zhou et al. 2018] but it also achieved good results in many other fields such as satellite imagery [Rakhlin et al. 2018; McGlinchy et al. 2019]. Basing on the architecture and on the pixel-based image segmentation, U-net has proven to be more successful than conventional models. This technique is also particularly effective with limited dataset images [Gadosey et al. 2020]. U-net is a symmetric encoder/decoder structure composed of a contracting and an up-sampling path (see Figure 2.1). In the contracting path the spatial information is reduced while the feature information is increased. In the expansive path the feature and spatial information are combined through a sequence of up-convolutions and concatenations with high-resolution features from the contracting path [Ronneberger et al. 2015]. Decreasing the spatial resolution allows the network to encode neighboring pixels into global information. The skip connections enables a combination of global and local information, in this

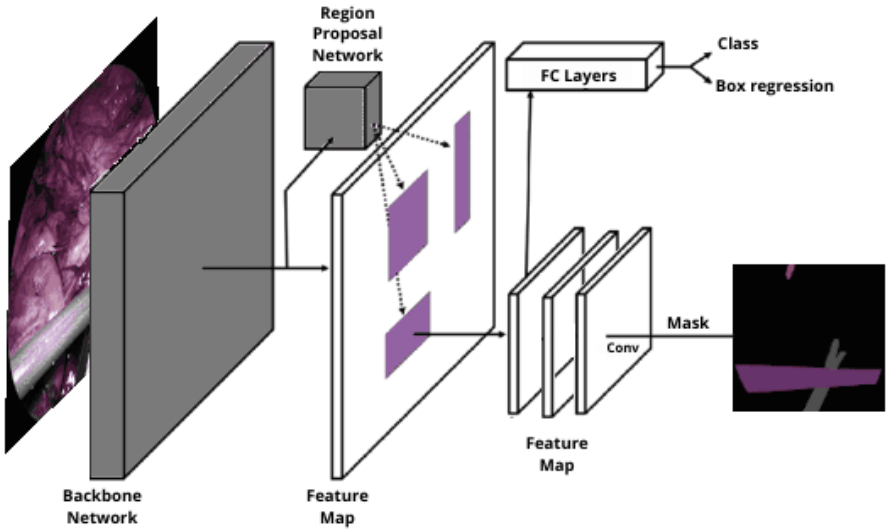


FIGURE 2.2: Network architecture of Mask R-CNN inspired by [K. He et al. 2017] . Low-level features are extracted from the input image and they are used by the Region Proposal Network (RPN) to produce Region of Interest (ROI). ROIs are processed in the network to predict bounding boxes, the class of each box and segmentation mask.

way details in the data that could get lost during the compression process can be preserved and localized precisely [Höhlein et al. 2020]. In the original architecture [Ronneberger et al. 2015] the contracting path consists of the repeated application of two 3×3 convolutions and a 2×2 maxpooling operation with stride 2 for downsampling. The expansive path consists of an upsampling of the feature map followed by a two 3×3 convolutions. In the final layer, a 1×1 convolution is used to map all 64 component feature vectors to the desired number of classes. All layers use ReLU [Dahl et al. 2013], except for the last layer, where Softmax [Gold et al. 1996] is used in order to select the best scoring category.

3.3 Mask R-CNN

Mask R-CNN [K. He et al. 2017] is an extension of Faster R-CNN [Ren et al. 2015] used to perform both object detection and instance segmentation. The authors introduced a new branch in Mask R-CNN to predict an object mask in parallel with the branch used for classification and bounding box regression. Explicitly, Mask R-CNN is composed of two main stages (see Figure 2.2): (1) **region proposal** which firstly generates proposals of the regions which might contain an object in the input image, (2) **instance segmentation/classification** which then predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on

the proposal computed in (1). Both stages are connected to the backbone structure.

Backbone is a standard CNN (typically, ResNet architecture [K. He et al. 2016]) that serves as a feature extractor. Precisely, the early layers is able to detect low-level features (edges and corners), and later layers can detect higher level features (e.g., surgical tools). Since the backbone is usually a pre-trained network, already been trained to detect multiple object (i.e., trained on ImageNet [Deng et al. 2009]), the performance of Mask R-CNN can be improved thanks to the previous learned features.

Region proposal is generated by a lightweight neural network called RPN which scans over the backbone feature map. In order to bind features to its raw image location, the network uses predefined bounding boxes with a certain locations and size (i.e., width, height, scale, and aspect ratio) as anchors. RPN relies on anchors to identify an object in the feature map and its bounding box size. In this way, RPN can reuse the extracted features efficiently and avoid duplicate calculations.

Instance segmentation/classification consists of lightweight CNNs which takes proposed regions from stage (1) and (i) assign them to several specific areas of a feature map level, (ii) predict the class of the object inside the bounding box and (iii) generates the corresponding segmentation mask of the object inside the bounding box [K. He et al. 2017]. At this stage, the ROI Align is used to map each individual proposed region onto the feature map and extract pixel values inside the region with the aim to reduce the unneccessary offsets in predicting mask and improve the accuracy.

4 Applications of deep learning in medical images

In this section, we will provide an overview of the latest advances in DL to various generic tasks in medical image analysis: segmentation, classification and detection.

Segmentation

During last years, many methods for biomedical image segmentation were based on U-Net architecture, due to its ability to segment images efficiently in various

kinds of vision tasks [C. Guo et al. 2020].

Wang et al. [B. Wang et al. 2019] proposed a novel approach based on two encoders, namely Dual Encoding U-Net (DEU-Net). A spatial path with large kernel was used to preserve the spatial information and a context path with multi-scale convolution block was defined to capture more semantic information. The approach has proven to be comparable w.r.t. the state-of-the-art methods on the digital retinal images for vessel extraction (DRIVE) [Staal et al. 2004], the child heart and health study (CHASEDB1) [Fraz et al. 2012] and structured analysis of retina (STARE) dataset [Hoover et al. 2000]. Similarly, Jha et al. [Jha et al. 2020] proposed a novel architecture namely Double U-Net to perform semantic image segmentation. The authors used two U-Net architecture in sequence, with two encoders and two decoders. The first encoder is VGG-19, trained on ImageNet. The approach outperformed U-Net and the baseline models in several experiments performed on the MICCAI 2015 segmentation challenge [Bernal et al. 2017], the CVC-ClinicDB [Bernal et al. 2015], the 2018 Data Science Bowl challenge [Codella et al. 2018], and the Lesion boundary segmentation datasets.

Another U-Net-based architecture was presented in [H. Huang et al. 2020]. The authors proposed a novel UNet 3+ composed of full-scale skip connections and deep supervisions, which incorporate low-level details with high-level semantics from feature maps in full scales with fewer parameter. The approach is designed to support the accurate segmentation especially for organs that appear at different scales in the medical image volume. The experimental results proposed by the authors have demonstrated that UNet 3+ outperformed the previous state-of-the-art approaches on both liver and spleen datasets [Christ 2017].

Classification

Zang et al. [J. Zhang et al. 2019b] proposed a synergic deep learning (SDL) model to perform biomedical images classification. They used multiple deep convolutional neural networks (DCNNs) simultaneously and the resulting learned image representations are concatenated as the input of a synergic network. The idea is to penalize misclassification: if one DCNN makes a correct classification, a mistake made by the other DCNN leads to a synergic error that serves as an extra force to update the model [J. Zhang et al. 2019b]. The authors showed that the proposed SDL model achieved the state-of-the-art performance using ImageCLEF-2015, ImageCLEF-2016,

ISIC-2016, and ISIC-2017 datasets. Zhang et al. [J. Zhang et al. 2019a] also proposed an attention residual learning convolutional neural network model for skin lesion classification in dermoscopy images, to overcome DCNNs limitations in classifying skin lesions due to the insufficiency of training data, inter-class similarity, intra-class variation. The approach achieved the state-of-the-art performance in the ISIC-skin 2017 dataset [Goyal 2019]

Sajjad et al. [Sajjad et al. 2019] proposed a novel CNN-based approach for multi-grade brain tumor classification. The authors first proposed a segmentation of tumor regions from an MR image and, secondly, they relied on extensive data augmentation to effectively train the proposed system, avoiding the lack of data problem. They performed a brain tumor grade classification using augmented, achieving promising results that are comparable to existing methods.

Detection

Xie et al. [H. Xie et al. 2019] proposed a pulmonary nodule detection framework with 2D CNN. The authors adjusted the structure of Faster R-CNN using two region proposal networks and a deconvolutional layer to detect nodule candidates. They trained three models for three kinds of slices for later result fusion through voting and, in particular, the misclassified samples are kept for retraining a model which boosts the sensitivity. The approach achieved a comparable results w.r.t. the state-of-the-art on LUNA16 dataset [Setio et al. 2017], with a sensitivity of pulmonary nodule candidate detection of 86.42%.

Guo et al. [L. Guo et al. 2020] proposed a real-time DL system for diagnosis of precancerous lesions and early esophageal squamous cell carcinomas. The model generates a probability heat map for each input of endoscopic images to indicate suspected areas. The approach achieved good results on both images and video dataset. Another real-time approach was presented by Liu et al. [Y. Liu et al. 2020] to surgical tool detection in Robot-Assisted Surgery. The authors proposed an anchor-free CNN architecture which models the surgical tool as a single point: the center point of its bounding box. The aim of the authors was to eliminate the need to design a set of anchor boxes, achieving 98.5%mAP and 100% mAP at 37.0 fps on the ATLAS Dione [Sarikaya et al. 2017] and Endovis Challenge [X. Du et al. 2018] datasets, respectively. In a similar contest, González et al. [González et al. 2020] proposed an Instance-based Surgical Instrument Segmentation Network (ISINet)

based on Mask R-CNN to perform semantic segmentation of surgical instruments in robotic-assisted surgery scene. The authors included a temporal consistency module that takes into account the previously overlooked and inherent temporal information of the problem [González et al. 2020]. The approach outperformed state-of-the-art methods in Endovis Challenge.

CHAPTER 3

Thesis topic

The main objective of this thesis is to exploit recent advances in AI-based technologies, especially DL techniques, to improve diagnosis classification and detection with the final aim of supporting healthcare providers in finding the most appropriate preventive interventions and therapeutic strategies. Such approaches represent powerful tools to analyze the complex mechanisms of several pathologies; however, the explainability of these approaches still remains a challenge, requiring careful validation and interpretation for human users. To address this task, we also propose a further investigation of Explainable Artificial Intelligence (XAI) to potentially ensure that the results obtained by AI methods are sound, correct, and justifiable in order to help in translating inferred knowledge into particular hypotheses that can be tested with real-life experiments.

We started investigating the ability of different DL approaches at solving image processing tasks. DL-based algorithms have shown a great deal of potential in extracting hidden patterns in high-dimensional data and discovering new biomarkers, showing considerable ability in learning without human supervision.

In addition, we explored the use of these techniques to automatically estimate diagnosis based on clinical and omics data in several medical context. A combination

of data reduction and data visualization techniques were defined to remove redundant irrelevant information and paving the way to the proper 2-D image-based representation of high-dimensional data, improving both prediction quality and computational efficiency.

Furthermore, we exploited proper techniques in order to build a system that allows for a partial opening of the DL-black box by means of proper investigations on the rationale behind the decisions; this can provide improved understandings into which pre-processing steps are crucial for better performance and which elements are most involved in the training process.

PART II

SUPPORTING BIOMEDICAL ANALYSIS USING DEEP LEARNING

CHAPTER 4

Using CNNs for Designing and Implementing an Automatic Vascular Segmentation Method of Biomedical Images.

Contents

1 Introduction

The assessment of vascular complexity in the lower limbs provides relevant information about peripheral artery diseases; in fact, vascular collaterals act as a sort of natural bypass system sustaining tissue perfusion downward of vascular occlusion [Prior et al. 2004]. Intuitively, they can exert a protective impact on limb ischemia, thus reducing symptoms and improving the outcome in patients with Peripheral Arterial Occlusive Disease (PAOD) [McDermott et al. 2011]. This is why assessing the collateral vascular network in patients with PAOD has a significant impact on both therapeutic decisions and on prognostic estimation.

Different clinical studies for the estimation of vascular collateral growth in PAOD patients resulted in conflicting results, mostly because of the technical difficulties in the quantification of vascular network and its flow capability. Such evaluation is currently carried out by human operators via visual inspection of cine-angiograms, resulting in scorings that are largely operator-dependent. The task is particularly hard because of the intrinsic difficulty, for human operators, in obtaining objective measurements; indeed, the operator has to memorize the anatomy in each frame and virtually reconstruct the whole vascular tree. This kind of approach is also prone to errors because: (i) the vascular system, consisting of vessels of highly variable shapes, irregularly fills a 3-D Euclidean space; (ii) the information comes in form of a set of images, each one covering a small subset of the entire area, while the scoring must refer to the whole anatomical district; (iii) the Field Of View (FOV) might include external objects, such as surgical instruments and tools, electrode cables, catheters and other anatomical structures, especially bone; (iv) the acquisition method (i.e. moving the C-arm or the patient couch) reduces the quality of images that are characterized of various lighting effects and motion blur.

In this work we define an automatic methodology for vessel tree extraction, with the goal of fostering more reliable clinical assessments in the described scenario.

Several studies investigated neural-network based methods for automatic vessel segmentation of medical images [S. Yang et al. 2018; Moccia et al. 2018], especially for retina segmentation [K. Hu et al. 2018; Alonso-Caneiro et al. 2018]; state-of-the-art solutions are capable to provide accurate segmentation of the structure of interest over static and well-contrasted images. Nevertheless, to the best of our knowledge, this is one of the first attempts to segment vessels in the ilio-femoral district on 2-D

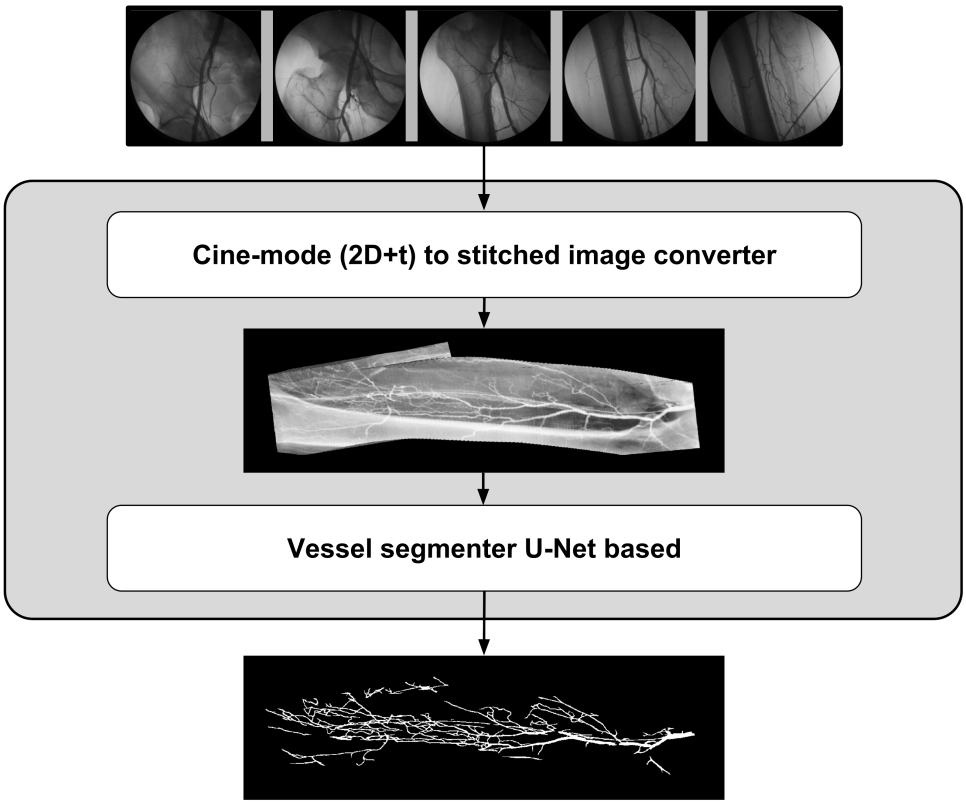


FIGURE 4.1: Workflow of the proposed framework.

projective images. In fact, the method presents several challenges: 1) non-trivial image pre-processing operations and feature detection are needed in order to elaborate and convert the cine-angiography into a whole static image for a larger FOV; ii) parameter fine-tuning of CNN layers is crucial to reach high segmentation accuracy. The remainder of the chapter is structured as follows. In Section 2 we provide a detailed description of our approach. Section 3 presents our experimental analysis performed on the results described in Section 3.1. Eventually, in Section 4 we draw our conclusion.

Part of the work proposed in this chapter has been published in [Bruno et al. 2018b].

2 Proposed Approach

The main goal of this work is to provide a new approach for automatic vessel segmentation from cine-angiography images. The workflow of the proposed framework, illustrated in Figure 4.1, can be divided into two steps: (i) a registration

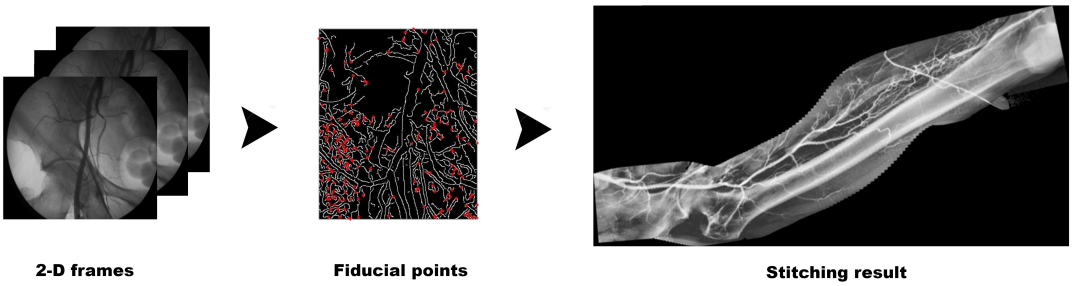


FIGURE 4.2: Exemplary images from cine-angiography to grayscale image. From left to right, the figure shows the original 2D frames over time, the fiducial points detected by feature extraction algorithms and the resulting stitching image.

method based on extracted feature used to built a static image with lager FOV, starting from different frames of cine-angiography frames and minimizing overlapping error, and (ii) an adaptation of U-net fully convolutional deep neural network architecture is used to segment the vascular tree from the stitched image.

In the following, we briefly describe the background techniques and methods, and provide further details on the proposed approach.

2.1 Cine-angiography stitching

In order to have a single image on which the vessel tree segmentation can be performed, the frames belonging to each cine-angiography are stitched together; as a result, a single static image inducing a larger field of view is obtained. To accomplish such task, in the preprocessing phase the 2-D images acquired over time are extracted from the video file (in DICOM format) and adaptive histogram equalization is performed on the negative of each image; then each frame t is geometrically aligned with the frame $t+1$ by matching corresponding fiducial points computed from each image. Speeded Up Robust Features (SURF) [Bay et al. 2006] and Maximally Stable Extremal Regions (MSER) [Matas et al. 2004] are used to identify the corresponding points to match. These features provide a sufficient number of points of interest that can be matched. For each consecutive frame, feature matching was performed by Random Sample Consensus (RANSAC) algorithm. The 2D roto-translation was computed by minimizing the Sum of Absolute Distance (SAD) between matching features.

This strategy is iterated over all the available frames and, finally, the stitched image is smoothed by a median convolution filter with a 3x3 kernel size.

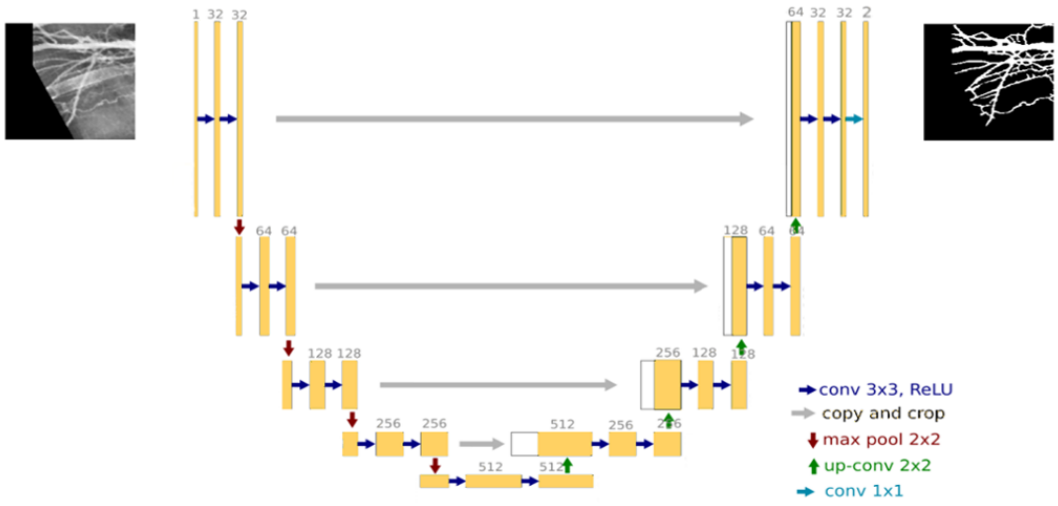


FIGURE 4.3: U-net architecture adapted from [Ronneberger et al. 2015].

In order to maximize the amount of fiducial points detected for each frame, features are computed on images preprocessed by means of the following operations:

1. negative image computation
2. adaptive histogram equalization
3. non-uniform illumination estimation
4. edge-detection using Canny filter [Canny 1987].

2.2 Network description

Vessel segmentation was performed by using U-net model described in Chapter 2, Section 3.2. In order to adapt the basic U-net architecture to the requirements of our problem, we introduced several modifications (see Figure 4.3):

1. the output segmentation layer is expanded from 1 to 2 feature maps to enable multi-class segmentation;
2. dropout [Srivastava et al. 2014] of 0.5 is used after each convolutional layer for addressing the overfitting problem;
3. batch normalization [Ioffe et al. 2015] is used in all layers to improve learning;

TABLE 4.1: Confusion matrix for vessel classification.

	Vessel present	**Vessel absent**
Vessel detected	True Positive (TP)	False Positive (FP)
Vessel not detected	False Negative (FN)	True Negative (TN)

4. the number of feature maps in all layers is reduced in order to accelerate deep network training.

The network has been trained with the categorical cross-entropy loss function [Kroese et al. 2013], which is defined as follow:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)]$$

where y_i are true labels, $\hat{y}_i$ are predicted labels and N is the number of classes.

3 Experimental Analysis

Rigorous performance evaluation of segmentation is an important step, since is not clear how to quantitatively evaluate a model. In the literature, segmentation performance is commonly evaluated with respect to ground truth manual segmentation performed by an expert clinician [Fenster et al. 2006]. In this study we evaluated our approach using two different quantitative methods: (i) Receiver Operating Characteristic (ROC) analysis [Bradley 1997]; (ii) Dice Similarity Coefficient (DCE) [Dice 1945].

According to Table 4.1, image pixels were labelled as TP, TN, FP and FN; sensitivity (S_e) and specificity (S_p) were computed as in Equation 4.1 and Equation 4.2, respectively.

$$S_e = \frac{TP}{TP + FN} \tag{4.1}$$

$$S_p = 1 - \frac{TN}{TN + FP} \tag{4.2}$$

The AUC of ROC evaluates the accuracy of pixel classification. An excellent model has AUC near to the 1 which means it has good measure of separability. In our

context, a higher AUC means that the model is able to distinguish between vessels and no-vessels, translating into very good segmentation.
The output of the network was converted in binary image to separate pixel values into two groups, black as background and white as vessels.

DCE measures the overlap between Ground Truth (GT) and Automatic Segmentation (AS) by means of the following formula:

$$2\frac{|GT \cap AS|}{|GT| + |AS|} \tag{4.3}$$

Result can range from 0 to 1, with 0 indicating no overlap and 1 indicating complete overlap [Crum et al. 2006].

Dataset. The proposed method was tested on a patient cohort including 30 cine-angiographies acquired at the Interventional Cardiology Units of Magna Graecia University Hospital (Catanzaro, Italy) and of Federico II University (Naples, Italy); patients gave explicit informed consent to the use of their anonymized data for research purpose. For each cine-angiography a stitched image was generated according to 2.1. The resulting images represent the entire ilio-femoral district for each patients. The vessel-to-background contrast is jeopardized by the presence of catheters, surgical tools, screws and other anatomical structures, especially bone; furthermore, the acquisition method reduces the quality of images that are characterized of various lighting effects and motion blur.

Ground truth segmentation for the input images were defined by expert clinicians. The entire dataset has been split in training set (80%) and testing (20%); in particular, the 20% of the training set was used as validation set in order to monitor the training process and prevent over-fitting. The test images arises from different patients than the train images. We also trained and evaluated U-net model using four-fold cross-validation (folds are split into disjoint sets of patients).

Training Phase. The input image and their corresponding segmentation maps are used to train the network. Since images are defined on a grayscale, the framework was used with the grayscale parameter enabled, so that only one input channel was used during the operation. The network was implemented in TensorFlow [Abadi et al. 2016] using the Keras wrapper and trained for 800 epochs, using the Adadelta optimizer [Zeiler 2012] with default parameters. In order to match the input shape of

TABLE 4.2: Results in terms of DCE for 12 subjects.

	DCE	**AUC**
CASE 0	0.591	0.984
CASE 1	0.645	0.991
CASE 2	0.646	0.986
CASE 3	0.633	0.992
CASE 4	0.726	0.997
CASE 5	0.627	0.992
CASE 6	0.561	0.974
CASE 7	0.648	0.987
CASE 8	0.737	0.992
CASE 9	0.494	0.990
CASE 10	0.686	0.986
CASE 11	0.601	0.987
MEAN	0.633	0.988
STANDARD DEVIATION	0.067	0.006

the network, the training/validation images are subdivided into 30600 tiles at a resolution of 128×128 pixels. For the network training we used only 24480 tiles (80%), while the remaining 6120 (20%) were used for validation. Tiles of single patient are not subdivided between training and validation set.

All experiments have been performed on a machine equipped with a 12 x86_64 Intel(R) Core(TM) CPUs @ 3.50GHz, running Linux Debian 7 and using CUDA compilation tools, release 7.5, *V*7.5.17 NVIDIA Corporation *GM*204 on GeForce GTX 970.

3.1 Results and Discussion

The experimental results from the test set are presented in Table 4.2. For each of the 12 subjects selected for testing on the given database, DCE and AUC are computed and are used as a measures of the performance. In order to improve visual inspection, a post-processing was applied with the aim of removing small objects of size less than 1×10^4 pixels w.r.t. the entire image from obtained results; this operation does not cause significant changes in terms of accuracy. The herein proposed approach achieves a DCE mean value of 0.633 ± 0.067. and a AUC mean value of 0.988 ± 0.006. It is worth noting that the high AUC mean value is influenced by the presence of background that is largely in our images. Then, we tried to minimize the background by cutting the edges of the images until white pixels appeared, since our class of interest is represented by the white pixels (i.e., the vessel tree). We also performed image binarization to select the most appropriate threshold, achieving an

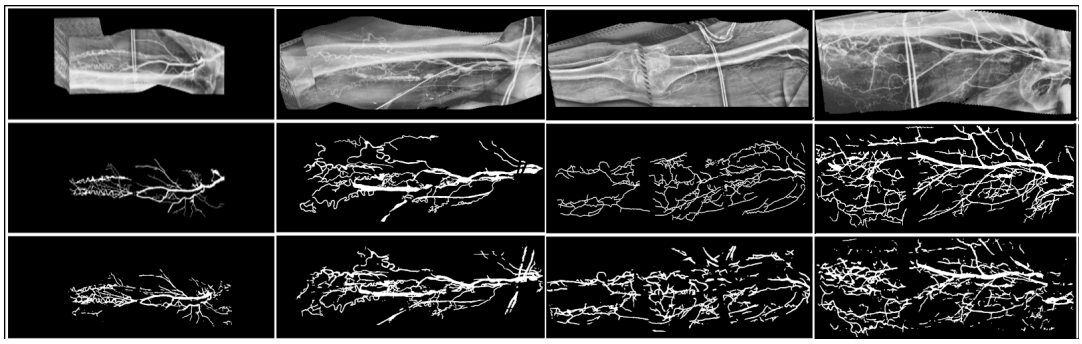


FIGURE 4.4: Results produced by U-net. The picture shows (from top to bottom) anatomical images, ground truth, automatic segmentations.

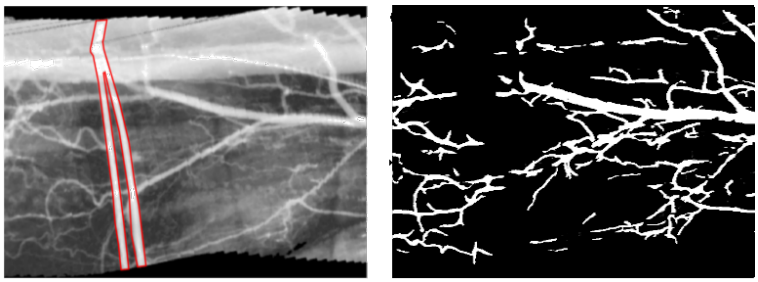


FIGURE 4.5: Example image featuring the presence of catheters (left) and the associated automatic segmentation (right). The red line circumscribes the catheters which appearance is very similar to the vessel (not marked in this image).

AUC mean $\pm$ std of 0.80 $\pm$ 0.01, and a min-max range of $0.70 - 0.90$.

Illustrative segmentation results, along with the manual segmentations and anatomical images, are shown in Figure 4.4. Results show that the network is able to distinguish between catheter and vessel with high precision, confirming the overall positive performance. In order to highlight the similarity between the vessels and the catheters and to show the correctness of automated segmentation, a more detailed view of the central case of Figure 4.4 is depicted in Figure 4.5.

Although very good ROC results are obtained, visual inspection shows some typical difficulties; in particular, false detection of noise and other artifacts near to the bones or in areas with excessive brightness. In order to tackle these issues, other pre-processing operations can be applied such as gamma correction that might help in a better enhancement of images [Kalyani et al. 2020]

This allows a potential interesting transfer-learning applications.

4 Conclusion

In this work we presented a novel ilio-femoral vessel segmentation approach based on the conversion of cine-angiographies into images with large FOV and the application of a properly improved U-net to detect more details and hard examples in the segmentation tasks. Experimental results show the effectiveness of our proposal, which is comparable to state-of-the-art methods.

As future work is concerned, we plan to improve the quality of the approach in order to obtain a segmentation which is even more close to manual segmentation; to this aim, a combination of multiple human-generated segmentations will be necessary in order to establish a ground truth and avoid human errors [Fritzsche et al. 2003].

Furthermore, we note that our approach currently takes into account only information local to each pixel through. One might think of including in the training phase useful information from shapes and structures present in the entire image. The automated vessel segmentation from ilio-femoral district provides the basis for automated assessment. Resulting segmentation images of the vessel pattern can be analyzed mathematically using nonlinear methods such as fractal [Novianto et al. 2003] analysis in order to provide numeric indicators of the extent of neovascularization. Experimental evaluations proved our proposal effective and robust, and this proposed automated vessel segmentation approach can be a suitable tool to be integrated into a complete prescreening system for PAOD detection.

CHAPTER 5

Multi-instance segmentation of medical instruments in endoscopic images

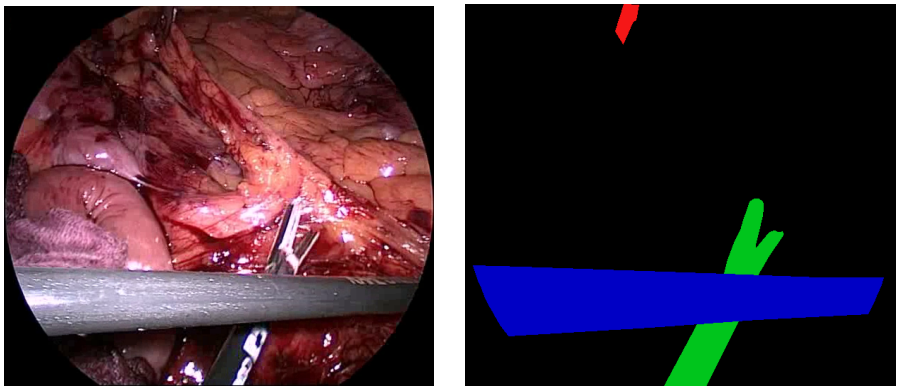
Contents

Disclosure:
The work presented in this chapter was carried out during a visiting period in the German Cancer Research Center (DKFZ - Heidelberg, Germany), in collaboration with Tobias Ross and under the supervision of Prof. Dr. Lena Maier-Hein, head of the Division of Computer Assisted Medical Interventions of the same institute.

1 Introduction

The number of minimally invasive surgeries (MIS) has increased continuously in the last decades thanks to significantly reduced hospitalization and recovery time, when compared to open surgery [Siddaiah-Subramanya et al. 2017]. However, since MIS rely on imaging the operating field through an endoscope, the complexity for the surgeon increases. MIS may affect surgeons' visual understanding due to limited field of view (FOV) [Bogdanova et al. 2016], lack of depth information [Bogdanova et al. 2016] and restricted freedom of movement [Azimian et al. 2010]. The surgical data science (SDS) community is working to develop novel methods and systems [Maier-Hein et al. 2017] with the aim of improving surgical vision and providing context-aware assistance. Main opportunities for SDS include: segmentation and classification [Singla et al. 2017], workflow modeling [S Swaroop Vedula et al. 2017] and context skill-assessment applications [Nguyen et al. 2019; Funke et al. 2019; S. Lin et al. 2019]. Often, context awareness applications rely on tracking of medical instruments [R. Wang et al. 2017; Shvets et al. 2018; Kletz et al. 2019], which can be either hardware [Qin et al. 2019] or video based [Bodenstedt et al. 2018].

The recent advancements in DL for computer vision enables achieving surgical instrument tracking relying from video-data only, providing a promising alternative to tracking systems [Bodenstedt et al. 2018; Allan et al. 2020; B. Lin et al. 2016]. However, two major challenges have to be tackled (see Figure 5.1), which are relevant to (1) endoscopic-image analysis (e.g., presence of noise, different levels of illumination, specular reflections, variations in background texture, presence of blood, smoke, blur) and (2) tracking instrument (e.g., unknown and varying number of distinct, small or overlapping instruments, miss-classification of tube-like structures as



a) RGB image frame b) multi-instance segmentation

FIGURE 5.1: (a) An example of video frame with two overlapping instruments and (b) its corresponding annotation.

instruments). While challenges in (1) have been widely handled by state-of-the-art methods, the tracking problem still remains an open problem [Dakua et al. 2019].

Multiple-instance segmentation represents a huge breakthrough in estimating the position of each instrument instance individually. While current video-based approaches have already achieved satisfactory performance in binary segmentation [Isensee et al. 2020], the topic of multi-instance segmentation received less attention. To the best of our knowledge, one of the first challenge on multi-instance instrument segmentation was presented by [Roß et al. 2020].

Through visual exploration of video frames [Maier-Hein et al. 2021], we experienced that the annotation of instrument instances could be improved by looking at previous frames. Based on this consideration, and on recent attempts in related research areas for video processing [Colleoni et al. 2019; Moccia et al. 2019b], we hypothesize that **(H1)** processing the temporal dynamics of consecutive frames may help in improving the identification of multiple-instances and their segmentation. Including temporal information may, however, be useless when instruments are not moving across consecutive frames. To take this into account, we hypothesize that **(H2)** including the probability of a pixel to be an instrument may further increase the robustness and generalizability of multiple instrument segmentation.

Guided by these research hypotheses, we make use of Mask R-CNN [K. He et al. 2017] which relies on (1) the raw video frame to be annotated, (2) the probability of a pixel to be an instrument and (3) the LSTM to summarize the optical flow estimation of instrument motion.

The remainder of the chapter is structured as follows. In Section 2 we illustrate

the related literature; in Section 3 we describe our approach to multiple-instance segmentation; in Section 4 we describe the dataset used, the evaluation metrics, the implementation details and the experimental activity performed. The results are presented and discussed in Section 5. Finally, we draw our conclusion in Section 6.

Part of this work has been submitted to *Medical Image Analysis* as: T. Roß, P. Bruno, A. Reinke, M. Wiesenfarth, L. Koeppel, P. M. Full, B. Pekdemir, P. Godau, D. Trofimova, F. Isensee, S. Moccia, F. Calimeri, B. P. Müller-Stich, A. Kopp-Schneider, L. Maier-Hein, "How can we learn (more) from challenges? A statistical approach to driving future algorithm development" [1], and is currently under revision.

2 Related work

In this section, we discuss the most relevant work on multiple instrument segmentation, including a detailed description of the effect of including temporal information for the multiple instance segmentation.

2.1 Multiple instance segmentation

Almost all recently published papers about multiple instance segmentation inside or outside the medical domain are based on a Mask R-CNN [Hafiz et al. 2020] which is a deep neural network aimed to solve instance segmentation and object detection (see Chapter 2 Section 3.3).

Mask R-CNN outperforms all existing methods in the MSCOCO [T.-Y. Lin et al. 2014] challenge, which consists in multiple objects segmenting in natural images. In the medical field, Mask R-CNN was first used for segmenting multiple instruments in videos of laparoscopic gynecology by Kletz et al. [Kletz et al. 2019]. However, the authors showed some limitations in identifying instruments when they are covered by another instrument or tissue. Indeed, especially in regions where instruments overlap or they are partial visible or covered by a tissue, the bounding box proposals can contain parts of multiple instrument instances, resulting in an accurate labeling and miss-segmentation.

[1] https://arxiv.org/abs/2106.09302

2.2 Temporal information processing

Allan et al. [Allan et al. 2015] first proposed to include temporal information through the use of optical flow, showing that using optical flow features improved the estimation of instrument pose and multi-class labeling. García-Peraza-Herrera et al. [Garcıéa-Peraza-Herrera et al. 2016] proposed a fully-convolutional neural network (FCNN) to perform instrument segmentation, based on the optical flow. They calculated the flow by identifying feature points (using the Shi-Tomasi corner detector) in consecutive frames. However, the detector requires that reflections and lightning conditions over consecutive frames are consistent. This assumption is hard to fulfill for laparoscopic videos. For estimating the pose of robotic instruments, Colleoni et al. [Colleoni et al. 2019] stacked temporally consecutive video frames to train a 3D FCNN. They assumed that in such a way the network will learn spatio-temporal features that help to localize the instrument joints. However, they did not further explore of how much temporal information should be included and their approach is used for segmenting instrument joints only. Jin et al. [Jin et al. 2019] were the first who proposed UnFlow, a CNN for the computation of optical flow [Meister et al. 2018]. They applied this approach to perform instruments segmentation and, specifically, they included optical flow to initialize the attention of a temporal attention pyramid network. Thus, their network learned to focus on moving objects. Furthermore, there is only little literature that explicitly addresses the problem of multi-instance instrument segmentation and neither Allan et al. [Allan et al. 2015], Colleoni et al. [Colleoni et al. 2019] nor Jin et al. [Jin et al. 2019] used the optical flow for the multi-instance segmentation.

In this chapter we propose multi-instance segmentation approach which relies on the use of temporal information and the high abilities of binary segmentation methods to even increase our algorithm performance.

3 Methods

Given a video $X = \{x_1, ..., x_{N_X}\}$, being N_X the total number of frames in X, the task is to segment all instrument instances $\{I_1, I_2, ..., I_{N_{classes}}\}$ in a frame x, at a discrete time step t. The proposed approach consists of three steps as shown in Figure 5.2:

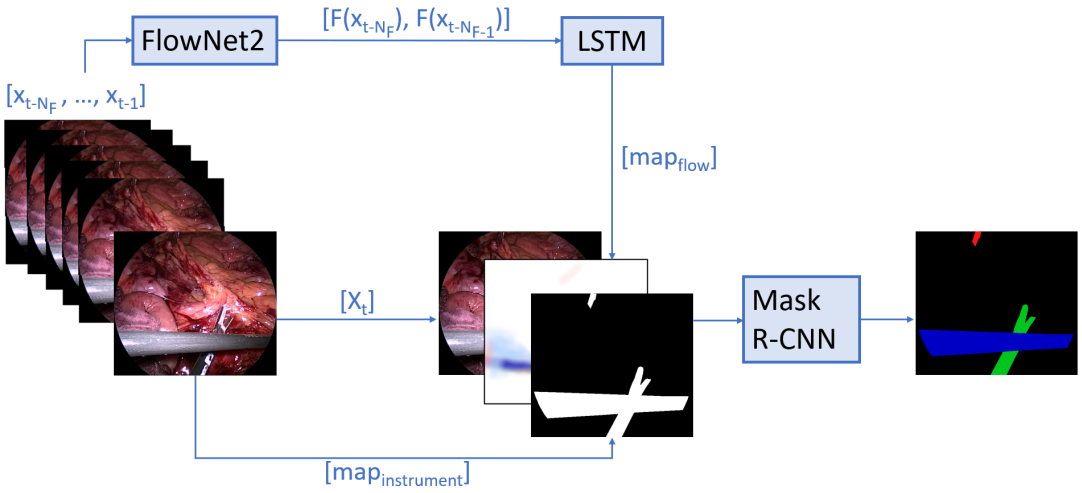


FIGURE 5.2: Workflow of the proposed approach to perform multi-instance segmentation. The $map_{instrument}$ is generated by an instrument segmentation network by using only the last frame x_t. FlowNet2 is used to predict the optical flow for each pair of frames $(F(x_{t-N_F}), F(x_{t-N_{F-1}}))$ and LSTM is used to summarize them in one image (map_{flow}). The Mask R-CNN is trained with the stacked input $[x_t, map_{instrument}, map_{flow}]$ to produce the instrument candidates.

1. Relying on frames in X, estimating the optical flow between two consecutive frames using FlowNet2 and LSTM to summarize the results of FlowNet2 computed over a video frames, resulting in an optical flow map, as described in Section 3.1.

2. For frame x_t, estimating the likelihood of a pixel to be an instrument, resulting in a instrument likelihood map, as described in Section 3.2.

3. Predicting of the multiple instances (instrument instance map) by concatenating the raw RGB frame, the instrument likelihood map and the optical flow prediction (step 1, step 2) as input for Mask R-CNN (see Section 3.3).

3.1 Prediction of optical flow

The core idea behind exploiting the optical flow is that identifying an instrument or separating two overlapping instruments is easier if their motion is observed over time. In fact, instruments usually move faster than background tissue. It is worth noting that the computation of optical flow relies on the assumption that the instruments' location changes over time. This problem may be attenuated by a preprocessing phase which evaluates if the optical flow changes between consecutive frames and, then, if the instruments are not static. If no motion is observed, the

pre-processing algorithm can discard the optical flow prediction for those specific frames and the multi-instance segmentation can be computed using only the RGB frame and the instrument likelihood map, reducing also the computation time. In this work, we first estimated the instrument motion using the optical flow concept for each pair of consecutive frames and, secondly, we summarized these movements into a single image to directly include the information of complete video sequence in the segmentation task, as reported in the following sections.

Optical flow estimation

The optical flow was estimated using a pretrained version of FlowNet2 [Ilg et al. 2017], which was trained on a large synthetic dataset named Flying Chairs [Dosovitskiy et al. 2019]. FlowNet2 harnesses the advantage of FlowNet [Alexey Dosovitskiy et al. 2015], a CNN that directly estimates the optical flow, resolving the problem with small displacements and noisy artifacts in estimating flow fields [Ilg et al. 2017].

The optical flow is computed over two consecutive frames (x_t and x_{t-1}) at time step t. The network takes these two consecutive frames and outputs a 2D-flow map of the size $[2 \times h \times w]$, where 2 represents the horizontal (F_x) and the vertical (F_y) component of the optical flow velocity and h, w the height and width of frame x_t.

LSTM-based summarization

In order to summarize the optical flow computed over all pairs of complete video sequence, we made use of LSTM [Hochreiter et al. 1997], an extension of recurrent neural network (RNN), which is used to find temporal-sequential patterns in time series data. The LSTM model is composed of the input layer, the recurrent layers, and the output layer. The recurrent layer is composed of memory block (set of recurrently connected subnets) instead of traditional neuron node; this memory block contains one or more self-connected memory cells and three multiplicative cells [Mou et al. 2019]: (i) forgetting gate that controls which information needs to be discarded from the past cell status, ignoring irrelevant features and automatically identifying the best input; (ii) input gate that determines the state used to update the unit; (iii) output gate that filters the output according to the state of the unit. The LSTM (1) takes in input the optical flow computed between all pairs of consecutive frames in a fixed sequence time, as described in previous section, and (2) outputs a

single map (map_{flow}) of size $[1 \times h \times w]$ which summarizes together all the optical flows received in input.

3.2 Prediction of instrument likelihood

Let x_t be a video frame, the objective is to predict the likelihood, for each pixel $i \in x$ where $i \in [0, 1]$, to be an instrument ($map_{instrument}$). The prediction is done with a 2D U-net [Ronneberger et al. 2015] inspired architecture, as described in Isensee et al. [Isensee et al. 2020]. The loss function ($\mathscr{L}_{comb}$) (Eq. 5.3) used to train the network is a combination of the regular pixel-wise cross-entropy loss ($\mathscr{L}_{CE}$) (Eq. 5.1) and the soft dice loss ($\mathscr{L}_{DCE}$) [Drozdzal et al. 2016] (Eq. 5.2). These loss function are defined as:

$$\mathscr{L}_{CE} = -\sum y_i \log(p_i) \tag{5.1}$$

$$\mathscr{L}_{DCE} = -\frac{2TP + \epsilon}{2TP + FP + FN + \epsilon} \tag{5.2}$$

$$\mathscr{L}_{comb} = \mathscr{L}_{DCE} + \mathscr{L}_{CE} \tag{5.3}$$

where y_i is the ground truth value and p_i is the estimated probability value for a given pixel i, TP (true positives), FN (false negatives), FP (false positives) are the number of instrument pixels correctly classified, incorrectly classified and the number of background pixels incorrectly classified, respectively.

3.3 Multiple instrument instance segmentation

Multiple instrument segmentation is achieved by a Mask R-CNN (see Chapter 2, Section 3.3) and, the input of the network is defined by stacking x_t, $map_{instrument}$ and map_{flow} in the channel dimension. We used Mask R-CNN since the number of visible instruments can vary, depending on surgery type and phases. The output of Mask R-CNN is a set of bounding box candidates with the corresponding predicted Intersection over Union (IoU) score and segmentation mask. Mask R-CNN generates a huge number of candidates in an image and the instances with a predicted $IoU < \delta$ are discarded, where δ represents the acceptance threshold (i.e., the instance is valid).

4 Experiments

4.1 Dataset

We evaluated our approach on the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge 2019 dataset [Roß et al. 2020; Maier-Hein et al. 2021], which consists of 10,040 endoscopic images that have been extracted from three different surgery types (rectal resection, proctocolectomy and sigma resection) with 10 videos each. The videos were recorded during daily routine at the University Hospital of Heidelberg (Germany). For each frame, a corresponding manually-segmented mask, containing all instrument instances exists. Specifically, dataset was split into 5,983 training and 4,057 testing images. The test set was split into three stages as follows:

- Stage 1: 663 cases where the data are taken from the same patients from which the training data were extracted
- Stage 2: 514 cases where the data are the same type of surgery as the training data but taken from patients not included in the training data
- Stage 3: 2,880 cases where the data are from a different but similar type of surgery

4.2 Metrics

For the evaluation, we used the multi-instance Dice Similarity Coefficient (DCE_{MI}) described in Equation 5.5 as defined in [Roß et al. 2020]. To calculate the DCE_{MI}, the DCE (see Chapter 4, Equation 4.3) is calculated between all predicted and ground-truth instrument instances for an image (see Equation 5.4). The Hungarian algorithm [Kuhn 1955] is also used to assign the correct DCE value for a ground-truth and prediction pair [Roß et al. 2020].

$$DCE(\hat{I}, I) = \frac{2TP}{2TP + FP + FN} \tag{5.4}$$

$$DCE_{MI} = \frac{1}{N_I} \sum_{i=1}^{N_I} DCE(\hat{I}_i, I_i) \tag{5.5}$$

where N_I is the number of instruments visible in the image, $\hat{I}_i$ the prediction and I_i the reference instance for instrument i.

4.3 Implementation details

Image flow prediction

We generated the optical flow with the FlowNet2 for the latest 5s of the video sequence resulting in flow map of the size $[n_{frames} \times h \times w]$. We summarized the optical flow using LSTM composed of 6 hidden convolutional layers of dimension $[32, 16, 8, 4, 2, 1]$.

Instrument likelihood prediction

The binary segmentation model is provided by Isensee et al. [Isensee et al. 2020] which was trained via 8-fold cross validation. Each model randomly selected patches of size 256x488 pixels for 2,000 epochs, with an epoch defined as an iteration of 100 batches. They relied on different data augmentation techniques such as image rotation, elastic deformation, scaling, mirroring, additive Gaussian noise, brightness, contrast and gamma augmentation [Isensee et al. 2020]. The model was trained using SGD as optimizer with a Nesterov momentum of 0.9 and a initial learning rate of 2^{-5} [Isensee et al. 2020].

Multi-instance segmentation

We used a ResNet-50 as backbone for the Mask R-CNN, with the anchor size of $(8, 16, 32, 64, 128, 256, 512, 1024)$ and aspect ratios of $(0.5, 1.0, 2.0)$. The detection threshold was 0.2 with the aim to also detect small instrument boundaries. Parameters where set according to small experiments during the implementation process. Training was conducted by using a batch size of 2 and SGD as optimizer.

4.4 Performance assessment

We investigated the effect of including (i) temporal information **(H1)** and (ii) instrument likelihood **(H2)**. We conducted an ablation study with the aim to compare the performance gains of the Mask R-CNN by combining or excluding **(H1)** and **(H2)** along with raw image as input of the network. We performed five experiments in total by concatenating the input as follows:

1. only raw images (we will refer in the following as $[R]$)
2. raw images and map_{flow} ($[R, F]$)
3. map_{flow} and $map_{instrument}$ ($[F, B]$)

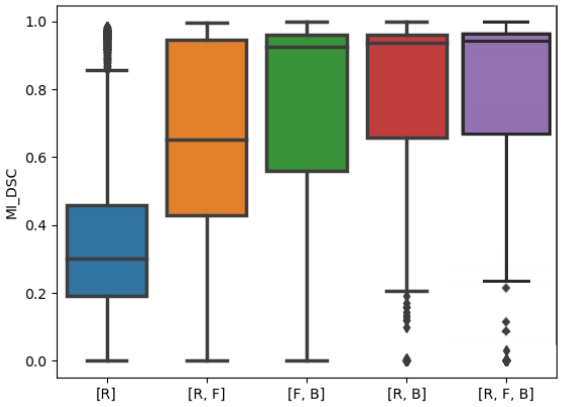


FIGURE 5.3: Boxplots of the performance of the models when trained on the raw only ($[R]$), on raw with the flow ($[R, F]$), on raw with likelihood ($[R, B]$), on flow with likelihood ($[F, B]$), and when combining all three ($[R, F, B]$).

TABLE 5.1: Performance results achieved by Mask R-CNN in terms of mean (μ), median ($\widetilde{x}$), 25th and 75th quartile (Q_1, Q_3), and the interquartile range (IQR) of the multi-instance dice coefficient DSC_{MI}. Mask R-CNN received as inputs the raw image ($[R]$), raw with optical flow ($[R, F]$), optical flow with instrument likelihood ($[F, B]$) and raw with the likelihood of an pixel being an instrument ($[R, B]$). The model $[R, F, B]$ is obtained using raw image, optical flow and instrument likelihood map.

Model	μ	$\widetilde{x}$	Q_1	Q_3	IQR
$[R]$	0.36	0.30	0.19	0.46	0.27
$[R, F]$	0.66	0.65	0.43	0.95	0.52
$[F, B]$	0.77	0.93	0.56	0.96	0.40
$[R, B]$	0.79	0.93	0.62	0.96	0.34
$[R, F, B]$	0.80	0.93	0.63	0.96	0.33

4. raw images and $map_{instrument}$ ($[R, B]$)
5. raw images, map_{flow} and $map_{instrument}$ ($[R, F, B]$)

For all experiments, we used the same hyperparameters in order to ensure the comparability.

5 Results and discussion

5.1 Ablation study

The results of all the tests performed in the ablation study are summarized in Table 5.1. It is worth to notice that using the same hyperparameters for all models, the performance of the model obtained using raw only as input ($[R]$) is much lower

than we expected; indeed, we obtained the worst result with the mean DSC_{MI} value of 0.36 and the median DSC_{MI} value of 0.30. We found that by incorporating the optical flow as input **(H1)** ($[R, F]$), the mean DSC_{MI} increased from 0.36 to 0.66 and the median DSC_{MI} increased from 0.30 to 0.65. On the other hand, the $map_{instrument}$ becomes very powerful when combined with raw images **(H2)** ($[R, B]$), increasing the mean DSC_{MI} from 0.36 to 0.79 and the median DSC_{MI} from 0.30 to 0.93. Concatenating the raw image, map_{flow} and $map_{instrument}$ ($[R, F, B]$) we achieved the best performance, by increasing the mean DSC_{MI} value from 0.79 to 0.80 and reducing the IQR from 0.34 to 0.33 w.r.t. ($[R, B]$).

Furthermore, after a visual inspection of the data, we noticed that images with crossing or close instruments are difficult to separate for our method as well as state of the art algorithms, resulting in mixed up classification masks [Roß et al. 2020]. This might be due to restrictions in the training and test data set, in fact, only the 36% of the data contains at least two and only the 8% of the images have more than two instrument instances. In addition, the instances overlap or intersect only in rare cases, reducing the possibility to train and to evaluate the separation capabilities of our algorithm.

6 Conclusion

In this work, we presented a DL-based approach to perform multiple instrument segmentation; we make use of a U-net to create a binary segmentation map and FlowNet2 and LSTM to estimate and summarize the optical flow, respectively. Both flow and binary segmentation were used to train a Mask R-CNN to predict the instrument instances. We hypothesized that by incorporating temporal information **(H1)** into the process in combination with the likelihood of an pixel to be an instrument **(H2)**, we can increase the performance of the approach. According to our experiments, we could show that including either the optical flow or the instrument likelihood the segmentation performance increased. The best performance was achieved by stacking the raw image with the optical flow and the instrument likelihood.

In conclusion, our approach is a good a step towards a satisfactory and robust multi-instance segmentation. As far as future work is concerned, we plan to improve the performance when the instruments overlaps using Conditional Random Field.

PART III

AI-based approach for Automatic Diagnosis using Gene Expression and Clinical Data

CHAPTER 6

Data Reduction and Data Visualization for Automatic Diagnosis using Gene Expression and Clinical Data

Contents

1 Introduction

Disease diagnosis is crucial for health condition monitoring [Health et al. 2007]. Data-driven disease classification can be extremely useful in medical diagnosis, which depends on a complex interaction of many clinical, biological and pathological variables. Notably, it is possible to discover relevant information and distinctive attribute related to specific diseases by properly analyzing different pieces of information coming from different "instances", such as gene expression or other clinical data.

Gene expression represent the amount of RNA produced in a cell under different biological states; such cell proportions are helpful to get a better insight into pathologies [J. Liu et al. 2017]; indeed, if the cells suffer from some disease (i.e., cancer or malignant tumors), genes are altered or mutated and, consequently, level proportions change. Clinical data provide both health-related information associated with patient care and features related to diseases that are relevant to diagnosis classification, and can even be useful for early prediction process. Both instances are composed by a set of "attributes" that characterizes each patient. The attributes can be defined as:

1. proportion of genes in each gene expression;
2. clinical information related to treatment, pathology and patient characteristics.

In terms of automated diagnosis, the massive amount of information provided by gene expression or clinical data might decrease the accuracy of the classifier. Hence, the set of available attributes should be reduced to a subset featuring the most significant variables (e.g., those capable of discriminating different classes). For this reason, feature selection is crucial to get rid of redundant and irrelevant features: it leads to a proper dataset dimensionality reduction, thus decreasing the complexity of classification and granting benefits in terms of reduced computational times and increased accuracy.

It is worth noting, indeed, that a dataset with a reduced set of features migh help to reduce overfitting issues and underline data by revealing interrelationships

among features [Tarek et al. 2017]. Furthermore, dimensionality reduction is often necessary in data visualization for helping analysts in studying the data, finding related observations and identifying relevant attributes associated with pathologies [Becht et al. 2019]. Given the importance of early and accurate medical diagnosis, especially for some specific pathologies, data visualization combined with ML techniques can be used to analyze high-dimensional multivariate data and automatically discover new biomarkers.

The main contribution of this work consists of a Data Visualization technique used to generate a set of 2-D images representing gene expression or clinical data of a set of patients with specific pathologies; in particular, we make use of heatmaps and hot-spot maps to show disease features proportion and their spatial correlation, respectively. Indeed, heatmap and hot-spot map provide two similar ways to display gene expression data, based on distribution and statistics, respectively. We performed experiments on the robustness of our method in comparison to current state of the art methods.

To the best of our knowledge, this is one of the first attempts at providing accurate diagnosis using data visualization maps as training factor.

Part of the work presented in this chapter has been published in [Bruno et al. 2020c].

2 Related work

Several methods were proposed to perform automatic diagnosis based on gene expression or clinical data. For instance, Support Vector Machines with Squared Loss (SVMSL) in [L. Zhang et al. 2018] were applied to perform cancer classification. The 1-norm SVMSL is first used to select "useful" genes and then to classify the resulting dataset. This method performs a very fast gene selection compared with other methods, but the classification performance is obtained using approximation methods that could affect the overall performance.

A DL-based multi-model ensemble approach was used in [Y. Xiao et al. 2018], where the authors first performed a differential gene expression analysis using five different classification models, namely K-nearest neighbors (kNN), SVMs, Decision Trees, Random Forests and Gradient Boosting Decision Trees, and then combined the

outputs via DL methods. Despite a high computational cost, the approach achieves accuracy over 98% in cancer classification.

Rotation Forest based-Genetic algorithms (GAs) were used to perform breast cancer classification based on clinical data [Aličković et al. 2017] featuring several attributes such as "clump thickness", "cell size", "cell shape", etc. GA is employed as a feature reduction mechanism to remove redundant data; the best feature subset is classified via Rotation Forest techniques. The approach achieves an accuracy mean value higher than 98%.

As for classification methods, DL-based models recently achieved promising results. DL models, such as CNNs [Esteva et al. 2017; Moccia et al. 2019a; Calimeri et al. 2019; Zaffino et al. 2019; Spadea et al. 2019], are proven to be appropriate and effective when compared to conventional methods. Interestingly, CNNs currently represent the most widely used method for image processing; nevertheless, the application of DL methods is not common in the context of gene expression, due to the well-known "large p, small n" problem [West et al. 2001], where p refers to the number of features and n refers to the number of samples. In the herein considered scenario, each sample is a 1-dimensional array of gene expression data or peculiar characteristics related to disease, which implies that traditional CNN models are not applicable for tumor classification [J. Liu et al. 2017]. A first attempt to deal with diagnosis using map-based CNNs is presented in [Bruno et al. 2019]. After data reduction, a set of heatmaps is generated and processed via a simple CNN; the approach achieves an accuracy mean value higher than 94%, but performance decreases while dealing with small datasets.

The remainder of the chapter is structured as follows. In Section 3 we describe the different type of data we worked on; in Section 4 we provide a detailed description of our approach, that has been assessed via a careful experimental activity, reported in Section 5; we analyze and discuss results in Section 6, eventually drawing our conclusions in Section 7.

3 Input

The input data herein considered for automatic diagnosis are:

1. Gene expression; it identifies the proportion of genes in a cell or tissue. Indeed, gene expression profiling is composed of genes strongly interrelate with each

others and it is also characterized by high dimension and noise; this makes many statistical methods inappropriate for the direct use on the expression data.

2. Clinical data; provide relevant pieces of information about patients (e.g., age, gender, education, etc.), treatments (e.g., kidney surgery, pump, oral medication) and disease characteristics (e.g., shape, density, size, nuclei, etc.). They are composed of a large number of attributes, each with their own data variability that could increase the complexity of classification [Bharti et al. 2014].

4 Proposed Approach

The herein proposed approach relies on the hypothesis that the use of data visualization technique might facilitate the task of disease classification. It consists of three steps, as illustrated in Figure 6.1:

1. Data Reduction, using PCA, in order to reduce the dataset dimension;
2. Data Visualization for each patient, relying on:
   - a heatmap, to visualize proportion level of each attibute;
   - a hot-spot map, based on Getis-Ord Gi*, to visualize spatial information level of each attibute;
3. Classification using CNNs.

In the following, we first describe Data Reduction technique in Section 4.2, then we illustrate Data Visualization technique in Section 4.3; eventually, we present the networks used to perform classification in Section 4.4

4.1 Data Pre-processing

We applied different pre-processing operations to transform raw data into a clean and tidy dataset and improve accuracy of analyses.
In particular, we performed (i) Data Cleaning steps, to deal with missing data, noise, outliers, and duplicate or incorrect records while minimizing the introduction of bias into the database; and (ii) Data Normalization, to remove systematic variation which affects the data.

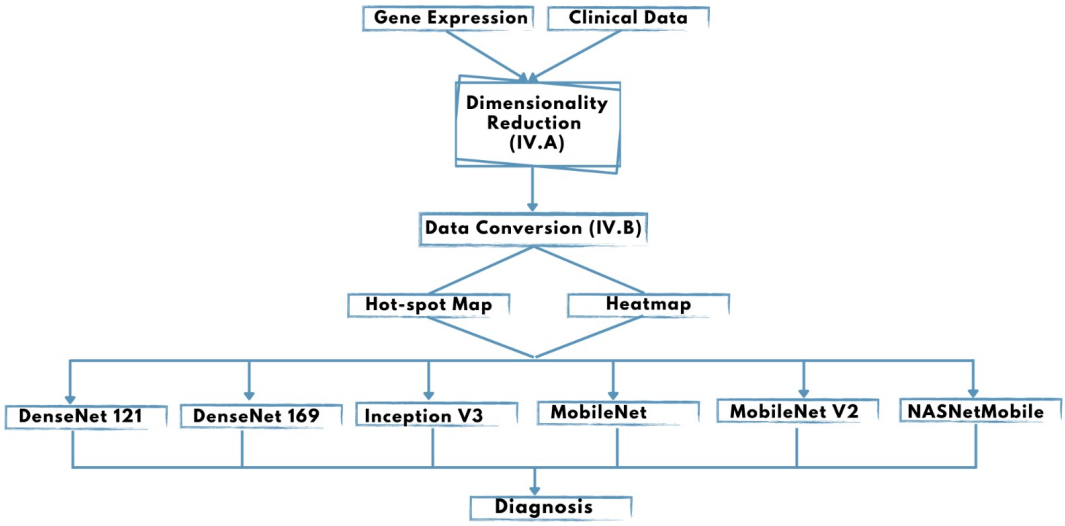


FIGURE 6.1: Workflow of the proposed framework. Dimensionality reduction is performed on Gene Expression or Clinical Data. The reducted dataset is trasformed into images and diagnosis classification is performed using 6 different CNNs.

TABLE 6.1: An example of gene expression of patients. Rows report small subsets of patients, while columns report first genes belonging to gene expression

PATIENTS	234002_AT	240140_S_AT	1557362_AT	221254_S_AT	221417_X_AT	240794_AT	...	STATUS
GSM38051	4.8	6.2	5.9	4.5	5.0	6.6		HEALTHY
GSM53062	10.9	14.6	11.0	7.9	7.5	26.6		ILL
GSM76577	14.6	38.8	14.0	11.3	12.1	17.4		ILL
...	...	...	...	...		...		

In our experiments, all the observations with at least one missing variable are discarded.

4.2 Data Reduction

PCA is used to reduce data from the high p-dimensional variable space to a lower K-dimensional variable component space, less than the number of samples n (see Chapert 1 Section 2.2). We selected a cut-off value of 75% of explained variance as it appeared to be the best compromise between data readability and quality.

4.3 Data Visualizzation

For each patient, we created heatmaps and hot-spot maps, as shown in Table 6.1 and Figure 6.2. Both maps are graphical representations of data, where the individual values contained in a matrix are represented as colors.

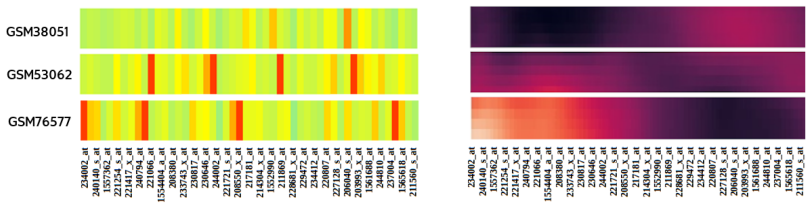


FIGURE 6.2: An example of data conversion: Heatmap (left) and hot-spot map (right) generated from data reported in Table 6.1. For each map, patients are represented along rows and genes along coloumns, respectively.

Heatmaps are directly generated from quantitative differences in expression levels, by coding numerical values of different types of data into colors. Hot-spot maps are generated using the Getis-Ord Gi* statistics computed to preserve the spatial information contained in the matrices; for example, in hotspot maps shows higher concentration of genes are represented by darker colors.

Getis-Ord Gi* statistic estimates the density distribution of features at the local level, and measures the degree of spatial association in a whole dataset. This kind of statistic serves as an indicator for local autocorrelation. The method evaluates the degree to which each feature (i.e., gene) is surrounded by features with similarly high or low values within a specified geographical distance (neighborhood).

The Getis-Ord *Gi** score of a feature i is calculated as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S \times \sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2\right]}{n-1}}} \tag{6.1}$$

where n is the total number of features, x_j is the attribute value associated with feature j and $w_{i,j}$ is the spatial weight matrix between points i and all neighbour points j (e.g. distance between points i and j). Furthermore:

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \qquad \bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \tag{6.2}$$

where S is the standard deviation of x_j and $\bar{X}$ is the mean of x_j. Typically, the spatial distances between observations at points were calculated by the Euclidean norm. The spatial weights $w_{i,j}$ were calculated using an exponential gravity model that is based on the distance $d_{i,j}$, i.e., the fractional distance between the features i and its neighbour j. In order to conform to Tobler's first law of geography (i.e., everything

is related to everything else, but closer things more so) [Sui 2004], a distance decay effect must be respected, and the value of the distance function must decrease when the distance increases. More formally, the partial derivative of the distance function with respect to distance should be negative, as follow $\frac{\delta w_{i,j}}{\delta d_{i,j}} < 0$.

The functions are combined with a distance cut-off criterion, such that $w_{i,j} = 0$ for $d_{i,j} > \alpha$. Then, let $d_{i,j} = 1$ the maximum pairwise distance, the spatial weights were calculated as follows:

$$w_{i,j} = e^{-\alpha \cdot d_{i,j}} \tag{6.3}$$

where α was set to 10, i.e., only the 10 nearest neighbors computed in the dataset were considered to calculate the spatial weight.

In our experiments, the output of the Getis-Ord Gi* statistic is a map indicating whether a feature belongs to a hot-spot (spatial cluster of high concentration of features), cold-spot (spatial cluster of low concentration of features) or an outlier (a high concentration of feature surrounded by low concentration of genes or vice-versa).

4.4 Classification

We compared the performances of six neural networks chosen on the basis of the good performance obtained on the *ImageNet* data set over several competitions [Rosebrock 2017]:

- DenseNet [G. Huang et al. 2017] is made of dense blocks, where for each layer the inputs are the feature maps of all the previous layers, providing more diversified features. This architecture is based on the observation that networks can be more efficient to train if they contain shorter connections between layers close to the input and those close to the output. The DenseNet architecture features several advantages, since each layer has direct access to the gradients from the loss function and the original input signal, the flow of information and gradients, ensuring alleviation of the vanishing gradient problem.

- Inception V3 [Szegedy et al. 2016] stacks 11 inception modules, where each module consists of pooling layers and convolutional filters with rectified linear units as activation function.

TABLE 6.2: Dataset description. For each dataset, on the columns we report the number of patients, the number of attributes, the number of attributes after PCA, the typologhy of dataset and classification.

DATASET	NUMBER OF PER PATIENTS	NUMBER OF ATTRIBUTES PER PATIENTS	NUMBER OF ATTRIBUTES PER PATIENTS AFTER PCA	ISTANCE	CLASSIFICATION
BREAST CANCER	569	10	8	CLINICAL DATA	BENIGN OR MALIGNANT MASS
MAMMOGRAPHY MASS	961	45	28	CLINICAL DATA	BENIGN OR MALIGNANT MASS
PARKINSON DISEASE	197	754	52	CLINICAL DATA	HEALTHY OR NOT
BREAST OR KIDNEY CANCER	635	9834	112	GENE EXPRESSION	BREAST OR KIDNEY
LYMPHOMA CANCER	62	4026	102	GENE EXPRESSION	HEALTHY OR NOT
BREAST CANCER (BC-TCGA)	590	17814	133	GENE EXPRESSION	HEALTHY OR NOT
BREAST CANCER (GSE2034)	286	12634	127	GENE EXPRESSION	HEALTHY OR NOT

- MobileNet [A. G. Howard et al. 2017] uses depthwise separable convolutions block; this significantly reduces the number of parameters. This block is composed of depthwise convolution layer that filters the input and pointwise convolution layer that combines these filtered values to create new features.
- MobileNet V2 [Sandler et al. 2018] introduces a new block called expansion layer; it expands the number of channels in the data before going into the depthwise convolution.
- NASNetMobile [Zoph et al. 2018] is composed by a set of blocks searched by reinforcement learning search method, not predefined by authors.

5 Experimental Protocol

5.1 Dataset description and training phase

For the experimental analysis we used the following datasets:

- the publicly available dataset from the Gene Expression Omnibus (GEO)[1], a database consisting of microarray, next generation sequencing (NGS) and other high-throughput data;
- the dataset from the UC Irvine Machine Learning Repository[2], a database consisting of heterogeneous set of features about different diseses and different categories of DNA microarray;

[1] https://www.ncbi.nlm.nih.gov/geo/
[2] https://archive.ics.uci.edu/ml/index.php

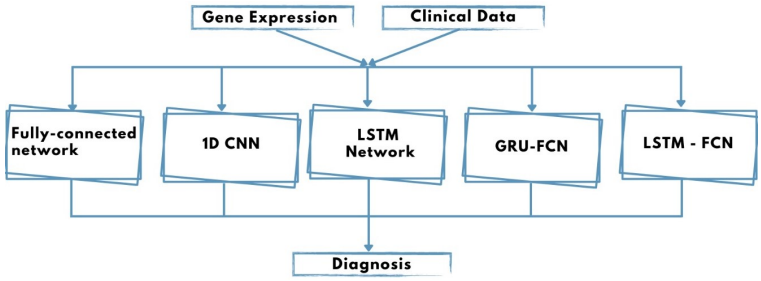


FIGURE 6.3: Workflow of *Approach C*: Gene Expression or Clinical Data is used to perform diagnosis classification using 5 different NNs.

- the dataset from The Cancer Genome Atlas (TCGA)[3], a publicly available database hosting many types of data including genomic, epigenomic, transcriptomic, and proteomic data.

All dataset used in our experiments are reported and described in Table 6.2. Each dataset is converted into 2-D images according to 4.3; we split them into training (80%) and testing (20%) sets; distribution of data across the two sets is identical w.r.t. all features. The 20% of the training set is used as validation set, in order to monitor the training process and prevent overfitting. Also, tiles of single patient are not subdivided between training and test set. In order to obtain a valid classification and to avoid majority class selection, we performed under-sampling to keep all sample in rare class and randomly select an equal number of samples in abundant class.

All experiments are performed on a machine equipped with a `12 x86 64` Intel(R) Core(TM) CPUs @3.50GHz, running GNU/Linux Debian 7 and using CUDA compilation tools, release 7.5, V 7.5.17 NVIDIA Corporation GM 204 on `GeForce GTX 970`.

5.2 Fine-tuning

For the training phase we performed hyperparameters optimization. Each network was trained with both optimizers Adam and SGD and for each optimizer 7 learning rate were tried. The best performance is obtained with the following configuration, trained for 150 epochs: Adam optimizer, learning rate 10^{-5} and batch size 32.

We also tried to modify the configuration of networks in terms of the number of nodes or levels to optimize the performance. We empirically acted changed the number of layers and we trimmed network size by pruning nodes to improve computational performance and identify those nodes which would not noticeably affect

[3] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

TABLE 6.3: Paired T-Test computed among heatmap and hot-spot map for clinical and gene expression data.

		P-VALUE
CLINICAL DATA	BREAST CANCER	0.009
	MAMMOGRAPHY MASS	0.011
	PARKINSON DISEASE	0.008
GENE EXPRESSION	BREAST OR KIDNEY	0.041
	LYMPHOMA CANCER	0.001
	BREAST CANCER (BC-TCGA)	0.030
	BREAST CANCER (GSE2034)	0.002

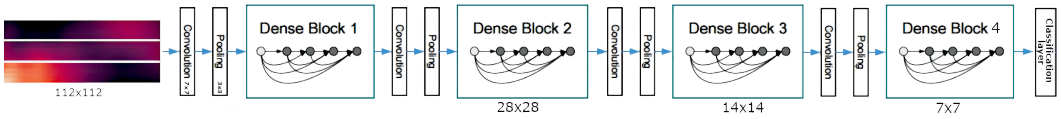


FIGURE 6.4: Architecture of the network DenseNet 169 inspired by [G. Huang et al. 2017].

network performance.

However, as one might have expected, since we performed the experiments using well-know networks already optimized, we achieved the best performance using the standard configuration as originally proposed by respective authors.

5.3 Test Description

We tested three different approaches to our method:

- *Approach A*: Data reduction is performed. Data visualization is performed. *Heatmaps* are used as input to the network.
- *Approach B*: Data reduction is performed. Data visualization is performed. *Hot-spot maps* are used as input to the network.
- *Approach C*: Gene expression data or clinical data are used as input of 5 networks, as shown in Figure 6.3. Given that the original data are composed of data-matrix, we selected Neural Networks that are different from the ones used in Approaches *A* and *B*, and more suited for the numerical data.

For the experimental analysis we used the following networks:

- Fully connected network [Hannun et al. 2019]; a mesh network in which each of the nodes is connected to every other node.
- One dimension convolutional neural network [Hannun et al. 2019]; it computes a weighted sum of the input channels or features along one dimension.

TABLE 6.4: Evaluation results (and standard deviation) for the 6 networks after 10-fold cross validation with the different inputs for each dataset used as input (DATASET). Breast Cancer (BC), Mammography Mass (MM), Parkinson Disease (PD), Breast or Kidney (BK), Lymphoma Cancer (LC), Breast Cancer (BC-TCGA), Breast Cancer (GSE2034) are the dataset used in our experimental analysis. Recall, Precision and F1_score are reported as R, P and F1, respectively. Highlighted results represent the most significant.

	DATASET	METRIC	DENSENET 121	DENSENET 169	INCEPTION V3	MOBILENET	MOBILENET V2	NASNET MOBILE
HEATMAPS	BC	R	0.97 ±0.03	**0.99** ±0.01	0.97 ±0.02	0.96 ±0.02	0.97 ±0.03	0.98 ±0.02
		P	0.97 ±0.02	**0.99** ±0.01	0.98 ±0.01	0.96 ±0.02	0.96 ±0.03	0.99 ±0.02
		F1	0.97 ±0.01	**0.99** ±0.01	0.97 ±0.01	0.96 ±0.02	0.96 ±0.02	0.98 ±0.02
	MM	R	0.85 ±0.06	0.86 ±0.05	**0.91** ±0.04	0.86 ±0.05	0.81 ±0.07	0.83 ±0.09
		P	0.87 ±0.06	0.87 ±0.04	**0.94** ±0.03	0.88 ±0.04	0.84 ±0.06	0.85 ±0.05
		F1	0.86 ±0.06	0.86 ±0.04	**0.92** ±0.03	0.87 ±0.04	0.82 ±0.06	0.84 ±0.08
	PD	R	0.98 ±0.02	**0.99** ±0.01	0.98 ±0.01	0.97 ±0.02	0.97 ±0.03	0.98 ±0.02
		P	**0.99** ±0.03	**0.99** ±0.01	0.98 ±0.02	0.98 ±0.01	0.98 ±0.02	0.98 ±0.01
		F1	0.98 ±0.02	**0.99** ±0.01	0.98 ±0.02	0.98 ±0.01	0.97 ±0.02	0.98 ±0.01
	BK	R	0.86 ±0.05	**0.91** ±0.04	0.82 ±0.06	0.82 ±0.06	0.82 ±0.06	0.87 ±0.05
		P	0.87 ±0.04	**0.93** ±0.03	0.83 ±0.04	0.82 ±0.06	0.83 ±0.05	0.88 ±0.03
		F1	0.86 ±0.04	**0.92** ±0.03	0.82 ±0.05	0.82 ±0.06	0.82 ±0.05	0.86 ±0.04
	LC	R	0.94 ±0.01	**0.97** ±0.02	0.95 ±0.02	0.97 ±0.02	0.96 ±0.02	0.90 ±0.04
		P	0.95 ±0.01	**0.99** ±0.01	0.97 ±0.02	0.98 ±0.01	0.97 ±0.02	0.93 ±0.02
		F1	0.94 ±0.01	**0.98** ±0.01	0.96 ±0.01	0.97 ±0.01	0.96 ±0.02	0.91 ±0.02
	BC-TCGA	R	0.98 ±0.01	**0.99** ±0.01	0.98 ±0.01	0.95 ±0.02	0.97 ±0.01	0.97 ±0.01
		P	**0.99** ±0.01	**0.99** ±0.01	0.98 ±0.01	0.97 ±0.01	0.98 ±0.01	0.98 ±0.01
		F1	0.98 ±0.01	**0.99** ±0.01	0.98 ±0.01	0.96 ±0.01	0.97 ±0.01	0.97 ±0.01
	GSE2034	R	0.89 ±0.03	**0.94** ±0.01	0.90 ±0.04	0.86 ±0.05	0.82 ±0.04	0.91 ±0.03
		P	0.91 ±0.02	**0.95** ±0.01	0.92 ±0.03	0.88 ±0.04	0.85 ±0.03	0.93 ±0.03
		F1	0.90 ±0.03	**0.94** ±0.01	0.91 ±0.03	0.87 ±0.04	0.83 ±0.03	0.92 ±0.03
HOT-SPOT MAPS	BC	R	0.97 ±0.03	**0.99** ±0.01	0.98 ±0.01	0.98 ±0.01	0.98 ±0.02	0.98 ±0.01
		P	0.99 ±0.02	**1.0** ±0.01	0.99 ±0.01	0.99 ±0.01	0.98 ±0.02	0.99 ±0.01
		F1	0.98 ±0.01	**0.99** ±0.01	0.98 ±0.01	0.98 ±0.01	0.98 ±0.01	0.98 ±0.01
	MM	R	0.90 ±0.03	0.90 ±0.02	0.92 ±0.03	**0.96** ±0.01	0.89 ±0.04	0.93 ±0.02
		P	0.92 ±0.02	0.92 ±0.02	0.96 ±0.01	**0.97** ±0.01	0.92 ±0.03	0.94 ±0.01
		F1	0.91 ±0.03	0.91 ±0.02	0.94 ±0.01	**0.96** ±0.01	0.91 ±0.03	0.93 ±0.02
	PD	R	0.98 ±0.02	**0.99** ±0.01	0.98 ±0.01	0.98 ±0.01	0.98 ±0.02	0.98 ±0.01
		P	0.99 ±0.01	**1.0** ±0.01	0.99 ±0.01	0.98 ±0.01	0.98 ±0.02	0.99 ±0.01
		F1	0.98 ±0.01	**0.99** ±0.01	0.98 ±0.01	0.98 ±0.01	0.98 ±0.02	0.98 ±0.01
	BK	R	0.89 ±0.04	**0.92** ±0.04	0.85 ±0.06	0.84 ±0.05	0.85 ±0.05	0.88 ±0.05
		P	0.91 ±0.03	**0.93** ±0.03	0.86 ±0.06	0.84 ±0.05	0.87 ±0.04	0.88 ±0.04
		F1	0.90 ±0.03	**0.92** ±0.03	0.85 ±0.06	0.84 ±0.05	0.86 ±0.04	0.88 ±0.04
	LC	R	0.96 ±0.02	**0.99** ±0.02	0.96 ±0.02	0.98 ±0.01	0.97 ±0.02	0.92 ±0.03
		P	0.97 ±0.02	**0.99** ±0.02	0.97 ±0.01	**0.99** ±0.01	0.98 ±0.02	0.93 ±0.03
		F1	0.96 ±0.02	**0.99** ±0.02	0.96 ±0.02	0.98 ±0.01	0.97 ±0.02	0.92 ±0.03
	BC-TCGA	R	0.99 ±0.00	**1.0** ±0.00	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01
		P	**1.0** ±0.00	**1.0** ±0.00	**1.0** ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01
		F1	0.99 ±0.00	**1.0** ±0.00	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01	0.99 ±0.01
	GSE2034	R	**0.94** ±0.03	0.93 ±0.02	**0.94** ±0.05	0.82 ±0.06	0.85 ±0.06	0.90 ±0.04
		P	**0.95** ±0.03	0.94 ±0.02	0.94 ±0.05	0.84 ±0.06	0.87 ±0.04	0.91 ±0.03
		F1	**0.94** ±0.03	0.93 ±0.02	**0.94**±0.05	0.83 ±0.06	0.86 ±0.05	0.90 ±0.03

- LSTM network [Hundman et al. 2018]; designed to avoid the long-term dependency problem. It is able to remember information for long periods of time, thanks to the chain structure.

- The GRU-FCN [Elsayed et al. 2018]; it combines the gated recurrent unit (GRU) with a temporal fully convolutional neural network (FCN). The FCN layers learn to extract feature representations from the data, without prior data reprocessing or feature engineering requirements. The GRU enables the model to recognize temporal dependencies within sequential data-streams. Therefore, the GRU-FCN model can learn both features and temporal dependencies to predict the correct class for each element in data training.

- The LSTM-FCN [Karim et al. 2017]; it combines a LSTM with a temporal fully convolutional neural network (FCN).

5.4 Performance Metrics

We assessed the effectiveness of our approach by means of three metrics: (i) Recall ($Rec = \frac{TruePositive}{TruePositive+FalseNegative}$), (ii) Precision ($Prec = \frac{TruePositive}{TruePositive+FalsePositive}$) and (iii) F1-score ($F1\_score = \frac{2*Prec*Rec}{(Prec+Rec)}$).

We perform paired t-test [Hsu et al. 2007] among heatmap and hot-spot map in order to check whether the population distributions are similar. The P-value of <0.05 was considered significant.

As shown in Table 6.3, p-value is less than the significant threshold; hence, the null hypothesis (i.e., there is no significant difference between the population means) can be rejected for all datasets.

In order to prove that the results are statistical different, we also computed paired t-test among the performance of all approaches (i.e., we compared *Approach A* against *Approach B*, then *A* against *C*, and finally *B* against *C*).

6 Results and Discussion

Table 6.4 reports classification results in terms of Recall, Precision and F1_score after 10-fold cross validation for all dataset, using either heatmaps or Getis-Ord Gi* hot-spot maps. Results show that the most efficient architecture was the DenseNet 169;

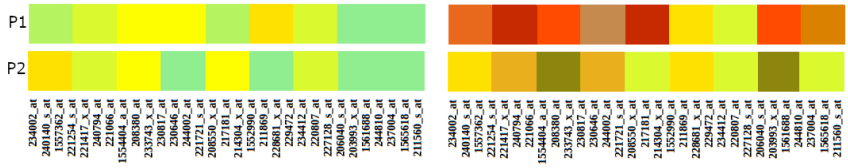


FIGURE 6.5: An example of heatmaps generated according to Breast/Kidney Data on left and to Lymphoma cancer on right. Rows data represent in the first row patient suffering from Breast Cancer (on left) and from Lymphoma (on right) and in second row another patient suffering from Kidney Cancer (on left) and healthy patient (on right).

hence, it was the one selected for the study; the specific layer composition used is depicted in Figure 6.4.

The herein proposed approach achieves promisingly results using both gene expression and clinical data. In general, we used Recall as our reference metric, since in this context the most important thing is to minimise False Negatives (i.e., disease is present but is not identified).

It considers prediction accuracy among only actual positives and explain how correct our prediction is among ill people.

For instance, taking into account gene expression data, the results of *Approach A* (i.e. heatmap) show the highest Recall mean value on Breast Cancer (BC-TCGA) (recall: 0.99 ± 0.01, precision: 0.99 ± 0.01, F1_score: 0.99 ± 0.01; among clinical datasets, instead, the highest Recall mean value is obtained on Parkinson Disease (recall: 0.99 ± 0.01, precision: 0.99 ± 0.01, F1_score: 0.99 ± 0.01 and Breast Cancer (recall: 0.99 ± 0.01, precision: 0.99 ± 0.01, F1_score: 0.99 ± 0.01. In general, due to the complexity of gene expression data and the similarity in the expression levels of each gene, the classification performance is better when we use clinical data for almost all datasets. Furthermore, the evaluation of the prognosis based on hot-spot map shows that the overall classification performance in terms of Recall is statistical significant better than using heatmap for almost all datasets.

As described in Table 6.3, heatmap and hot-spot map have different population distribution. This means that the spatial information generated by Getis-Ord Gi* statistic is critical for providing knowledge and improving both precise population information and classification accuracy. Considering the gene expression data, since Getis-Ord Gi* assesses whether a significant dependency exists between the spatial distributions of points (i.e., genes) and their associated marks (i.e., expression levels),

TABLE 6.5: Evaluation results (and standard deviation) for the 5 neural networks after 10-fold cross validation for each dataset. Highlighted results represent the most significant. Recall, Precision and F1_score are reported as R, P and F1, respectively.

Dataset	Metric	Fully connected	CNN 1-D	LSTM	LSTM FCN	GRU FCN
Breast Cancer	R	0.94±0.02	0.94±0.02	0.96±0.02	0.92±0.03	**0.96**±0.03
	P	0.95±0.02	0.94±0.02	0.96±0.02	0.93±0.03	**0.97**±0.02
	F1	0.94±0.02	0.94±0.02	0.96±0.02	0.92±0.03	**0.96**±0.02
Mammography Mass	R	**0.84**±0.04	0.83±0.05	0.76±0.03	0.81±0.02	**0.84**±0.02
	P	0.84±0.04	0.83±0.05	0.78±0.03	0.82±0.02	**0.85**±0.02
	F1	0.84±0.04	0.83±0.05	0.77±0.03	0.81±0.02	**0.84**±0.02
Parkinson Disease	R	0.94±0.02	**0.95**±0.02	0.93±0.02	**0.95**±0.02	0.94±0.02
	P	0.95±0.02	**0.96**±0.02	0.94±0.02	0.95±0.02	0.95±0.02
	F1	0.94±0.02	**0.95**±0.02	0.93±0.02	0.95±0.02	0.94±0.02
Breast or Kidney	R	0.84±0.05	0.82±0.06	0.82±0.06	0.81±0.06	**0.86**±0.03
	P	0.85±0.04	0.82±0.06	0.83±0.05	0.82±0.06	**0.87**±0.03
	F1	0.84±0.04	0.82±0.06	0.82±0.05	0.81±0.06	**0.86**±0.03
Lymphoma Cancer	R	0.91±0.01	0.93±0.02	0.93±0.02	0.94±0.02	**0.95**±0.02
	P	0.92±0.01	0.93±0.02	0.94±0.02	**0.96**±0.02	**0.96**±0.02
	F1	0.91±0.01	0.93±0.02	0.93±0.02	0.95±0.02	**0.95**±0.02
Breast Cancer(BC-TCGA)	R	0.92±0.02	0.92±0.02	0.89±0.04	0.91±0.03	**0.92**±0.01
	P	0.92±0.02	0.93±0.02	0.90±0.03	0.92±0.02	**0.93**±0.01
	F1	0.92±0.02	0.92±0.02	0.89±0.03	0.91±0.02	**0.92**±0.01
Breast Cancer(GSE2034)	R	0.60±0.05	0.61±0.08	0.62±0.07	**0.67**±0.08	0.65±0.06
	P	0.61±0.05	0.62±0.07	0.62±0.06	**0.67**±0.08	0.66±0.06
	F1	0.60±0.05	0.61±0.07	0.62±0.06	**0.67**±0.08	0.65±0.06

TABLE 6.6: Results in terms of Recall (R), Precision (P) and F1_score (F1) mean value of the best CNN model (DenseNet169) combined to hot-spot map (I) and the best ANN model (GRU-FCN) (II) for each dataset.

Dataset	Metric	DenseNet169	GRU-FCN
Breast Cancer	R	0.99±0.01	0.96±0.03
	P	1.0±0.01	0.97±0.02
	F1	0.99±0.01	0.96±0.02
Mammography Mass	R	0.90±0.02	0.84±0.02
	P	0.92±0.02	0.85±0.02
	F1	0.91±0.02	0.84±0.02
Parkinson Disease	R	0.99±0.01	0.94±0.02
	P	1.0±0.01	0.95±0.02
	F1	0.99±0.01	0.94±0.02
Breast or Kidney	R	0.92±0.04	0.86±0.03
	P	0.93±0.03	0.87±0.03
	F1	0.92±0.03	0.86±0.03
Lymphoma Cancer	R	0.99±0.02	0.95±0.02
	P	0.99±0.02	0.96±0.02
	F1	0.99±0.02	0.95±0.02
Breast Cancer(BC-TCGA)	R	1.0±0.00	0.92±0.01
	P	1.0±0.00	0.93±0.01
	F1	1.0±0.00	0.92±0.01
Breast Cancer(GSE2034)	R	0.93±0.02	0.65±0.06
	P	0.94±0.02	0.66±0.06
	F1	0.93±0.02	0.65±0.06

hot-spot maps could be successfully able to reveal insights into the genes organizations of tissue and organs. The same applies to clinical information since Getis-Ord Gi* could be used as pilot directions for the spatial analysis of any disease taking advantage of clinical, treatment and disease information for understanding the development and evolution of a disease. Thus, in pixel-based classification processes, spatial information turns out to be significant for the estimation of risk elements and peculiarities using gene expression or clinical information.

In addition, results also suggest that, in general, the network performance decreases when the expression levels or clinical data proportions are similar among different diseases. This is not surprising; for instance, a visual inspection of the heatmaps generated on Breast or Kidney Data (recall: 0.91 ± 0.04, precision: 0.93 ± 0.03, F1_score: 0.92 ± 0.03 using Heatmap and recall: 0.92 ± 0.04, precision: 0.93 ± 0.03, F1_score: 0.92 ± 0.03 using Hot-spot map) reveals that the proportion levels of genes seem to be similar in the diseases, as shown in Figure 6.5 on left. As a comparison, Figure 6.5 on right shows more evident difference proportion levels of genes in Lymphoma Cancer.

TABLE 6.7: Classification results compared to state of the art method, showing the results of DenseNet169 (our best CNN model), kNN and SVMs in terms of Recall (R), Precision (P) and F1_score (F1) mean value (and standard deviation) for each dataset. Highlighted results represent the most significant.

Dataset	Metric	DenseNet169	kNN	SVMs
Breast Cancer	R	**0.99**±0.01	0.96±0.03	0.97±0.02
	P	**1.0**±0.01	0.97±0.03	0.97±0.02
	F1	**0.99**±0.01	0.97±0.03	0.97±0.02
Mammography Mass	R	**0.90**±0.02	0.70±0.16	0.79±0.10
	P	**0.92**±0.02	0.72±0.10	0.81±0.10
	F1	**0.91**±0.02	0.71±0.11	0.80±0.10
Parkinson Disease	R	**0.99**±0.01	0.88±0.01	0.92±0.04
	P	**1.0**±0.01	0.89±0.01	0.94±0.03
	F1	**0.99**±0.01	0.88±0.01	0.93±0.03
Breast or Kidney	R	**0.92**±0.04	0.89±0.10	0.90±0.06
	P	**0.93**±0.03	0.90±0.09	0.91±0.06
	F1	**0.92**±0.03	0.89±0.09	0.90±0.06
Lymphoma Cancer	R	**0.99**±0.02	0.96±0.09	**0.99**±0.01
	P	**0.99**±0.02	0.97±0.09	**0.99**±0.01
	F1	**0.99**±0.02	0.96±0.09	**0.99**±0.01
Breast Cancer(BC-TCGA)	R	**1.0**±0.00	0.90±0.02	0.98±0.01
	P	**1.0**±0.00	0.92±0.02	0.98±0.01
	F1	**1.0**±0.00	0.91±0.02	0.98±0.01
Breast Cancer(GSE2034)	R	**0.93**±0.02	0.75±0.08	0.86±0.01
	P	**0.94**±0.02	0.77±0.07	0.87±0.01
	F1	**0.93**±0.02	0.76±0.7	0.86±0.01

This suggests that the proportion level of genes is too homogeneous, in the two pathologies, to allow an easy diversification in gene expression.

As discussed in Section 5, we performed classification tasks on the original, non-manipulated datasets (*Approach C*) in order to prove the effectiveness of the herein proposed method; in particular, according to what introduced in Section 5.3, we tested 5 different Neural Networks. The configuration for each network is obtained after 10-fold cross-validation used for hyperparameter tuning, in order to choose the parameter value that gives the lowest cross-validation average error; experiments were performed on the very same machine with the same configuration of the other approaches. Results in terms of Recall, Precision and F1_score are shown in Table 6.5. In the end, GRU-FCN obtained the best results over almost all datasets; for this reason, we select it as a yardstick for assessing our map-based approach, as described in Table 6.6.

Results show that performances of hot-spot map-based approach are definitely higher than the ones of a GRU-FCN-based system. Indeed, using DenseNet 169 we

obtain a Recall mean value higher than what resulting after GRU-FCN application, over all datasets used for our experiments. DenseNet 169 outperforms GRU-FCN on Breast/Kidney Data (i.e., recall: 0.92 ± 0.04 and recall: 0.86 ± 0.03 respectively); this is considered, as said before, the "most hard" dataset, due to similarities among the pathologies. In particular, performances are significantly improved on the Breast Cancer dataset (GSE2034): Recall value does not exceed 0.65 using GRU-FCN, while our approach reaches 0.93.

We computed paired t-test among the performance of the two approaches in order to prove that the results are significantly different. Our approach, based on CNNs, is conceived in order to overcome the problem by using a combination of PCA and visual representation (i.e., heatmaps and hot-spot maps): this helps at highlighting correlations among variables and quantitative differences or disparities in expression levels.

6.1 Comparison to the state-of-the-art

In order to assess the validity of the herein presented approach, we compared our algorithm to the most widely used state-of-the art methods: K-nearest neighbors (kNN) and Support-Vector Machines (SVMs). We compared our algorithm in terms of Recall and we performed a t-test for each algorithm pair to test for statistical significance.

Interestingly, the classification results of our approach are better (p-value < 0.05) w.r.t. kNN and SVMs for almost all datasets, as shown in Table 6.7. Only using the Lymphoma Cancer dataset classification based on SVMs (recall: 0.99 ± 0.01, precision: 0.99 ± 0.01, F1_score: 0.99 ± 0.01) there is no statistical difference (recall: 0.99 ± 0.02, precision: 0.99 ± 0.02, F1_score: 0.99 ± 0.02). In general, our image-based approach improves the evaluation of the prognosis. This result suggests that using data reduction and visual processing we are able to identify the most relevant features of the disease more than a data-based approach.

7 Conclusion

In this work we presented a novel approach to estimate diagnosis in several medical contexts; it is based on the conversion, or transformation, of selected pieces of information within different types of data into 2-D images. The use of images for

representing data presents two relevant advantages: first of all, it significantly improves data visualization; furthermore, it allows to take advantage of techniques that are explicitly geared towards image processing in order to perform classification tasks. In particular, starting from numerical "raw" data, we make use of PCA to reduce the dimension getting rid of redundant or irrelevant information and paving the way to the proper 2-D image-based representation.

We fine-tuned the approach by means of accurate experimental activities; in particular, we tested two different visual representations, namely heatmaps and hotspot maps, and six different CNNs for the classification. Experimental results show that not only our proposal is comparable to current state-of-the-art methods (i.e., kNN and SVMs), proving to be effective and robust, but also that it outperforms state-of-the-art non-convolutional NN-based approaches.

Classification accuracy is higher when the visual representation is based on hotspot maps rather than heatmaps; this suggests the importance of spatial information in both target detection and classification for each dataset used in our experiments. Accuracy is also higher when the classification task is about discriminate between presence and absence of a specific disease rather than about distinguishing among different pathologies. Indeed, given that the organisms rapidly adjusts their transcription patterns in response to environmental conditions or health status, the variety of gene expression characteristics or clinical information changes in response to the occurrence of a specific disease. This might result in an easy definition of patterns that identify these changes with respect to normal conditions. On the contrary, patients suffering from different diseases may have similar variations in gene expression or clinical information, and this condition may lead to lower classification accuracy. In order to tackle this issue, we believe that a quantitative assessment of the similarities among diseases or tissues of origin with reference to the corresponding gene expressions or clinical information of each patient should be considered [Sandberg et al. 2005].

In this context, quality assessment of microarray data is an important and often challenging aspect, especially in gene expression analysis, which frequently involves the examination of a variety of summary statistics and diagnostic plots. In particular, data quality may be defined in terms of accuracy (systematic bias between the true and measured value), precision (the uncertainty in replicated measures) and it can suffer from outliers, unbalanced classes and missing values [B. E. Howard et al.

2009]. In order to properly address the problem, it is necessary to perform a thorough quality control check on that data. For instance, by analyzing the data it is possible to identify samples with reduced quality or quantity of proportion levels which will either fail to be detected in classification task (false negatives) or will have higher-than-expected counts (false positives); in order to improve classification, it would be necessary to discard these data [Mcdade et al. 2015].

In general, as one might expect, dataset quality as well as a careful dimensionality reduction and feature selection schemes are of great importance for effective CNNs, and subsequently for accurate disease predictions. As usual, choosing the most informative and important feature subset for training a model results in robust and competitive models.

As future work is concerned, we plan to develop a tool to offers useful hints for automatic diagnosis, in order to support clinicians in the actual clinical practice. With this aim, and to further improve the quality of the approach, we plan to analyze and identify major regions and features involved in the classification; in particular, we think of using Gradient-weighted Class Activation Mapping (GradCAM) [Selvaraju et al. 2017]. This approach has been widely used in the context of image classification to provide an explainability of DL-based models, allowing the definition of localization map able to describe the important regions in the images used to achieve a specific outcome [Selvaraju et al. 2017]. In our context, GradCAM can be used to highlighting the most relevant regions in the heatmaps used as training input; these regions could help in interpreting the network decisions and, pontentially, in discovering features (e.g., genes) related to the disease.

CHAPTER 7

Classification and survival prediction in Diffuse Large B-cell Lymphoma by gene expression profiling

Contents

1 Introduction

The univocal identification of cancer and the understanding of its composition are crucial in medicine; however, they represent non-trivial challenges. In order to extrapolate features of single cells from complex tumor admixtures, non-trivial approaches and accurate statistics analyses are required.

Among the different kinds of cancer, the treatment of Lymphoma requires some of the most difficult tasks; indeed, a proper understanding the conditions in which it arises is still an open problem, as well as the definition of the specific kind of genetic mutation causing its growth [Health et al. 2007]. Furthermore, we know that DNA changes related to Lymphoma are usually acquired after birth, rather than being inherited [Sandlund et al. 2016]; nevertheless, even if they may result from several causes, such as exposure to radiation, cancer-causing chemicals or infections, changes occur for no apparent reason, in general.

In order to effectively tackle these challenges, new techniques have been recently developed that enhance already existing immune profiling technology [Zhao et al. 2010]. In this context, statistical analysis of gene expression [Van't Veer et al. 2002] plays a crucial role, and it can be of help for immune profiling, therapeutic design, treatment strategies and also for studying and understanding the unusual growth and/or the migration of cells into organs or tissues from their sources of origin. For instance, in malignant tumors, levels of cellular infiltration are associated with tumor growth, cancer progression and patient outcome.

Several methods for prognosis prediction of Diffuse Large B-cell Lymphoma and analysis of gene expression profiling have been proposed, based on Fuzzy Neural Networks [Ando et al. 2002], statistical techniques [Dabney 2005], survival analysis [Hedström et al. 2015] or microarray manipulation [Khoshhali et al. 2012], among others. In recent years, a new research trend has been arising, mostly based on discover Cell-of-Origin (COO) into two distinct molecular subtypes, identified by gene expression profiling: the activated B-cell-like (ABC) and the germinal center B-cell-like (GCB)[Lenz 2013]. Indeed, the assignment of Diffuse Large B-cell Lymphoma into COO groups has become increasingly important with the emergence of novel therapies that have selective biological activity in GBC or ABC groups [Scott et al. 2014]. Many studies take advantage from different feature extraction methods to discover independent components from gene expression profile, such as PCA [Shulin

Wang et al. 2006], Linear Discriminant Analysis [Sharma et al. 2008] and Locally Linear Discriminant Embedding [B. Li et al. 2010], and Prediction Analysis for Microarrays (PAM) [Orsborne et al. 2011]. Although such methods have solid biomedical support, there are a great number of gene subsets with the same predictive performance which could lead to the arbitrariness selection of candidate gene subsets. In fact, each method suffers from some drawbacks, and many factors such as normalization, small sample size, noisy data, improper evaluation methods, and too many model parameters can lead to the overfitting of the resulting models, the bias of results and even false discovery [S.-L. Wang et al. 2013]. Among these methods, promising results has been attained by the work of Dabney *et al.* [Dabney 2005], that showed Classification to Nearest Centroids (ClaNC) outperforming other methods in terms of accuracy and overall error.

In this work we propose a novel approach for class prediction from gene expression data and survival analysis of tumor and immune subtype. The goal of this study is to find a correlation between genes and the disease. Specifically, the approach is based on the analysis of the survival probability trend on two groups generated according to a ML approach called linear Support Vector Regression [Awad et al. 2015]. The implementation herein presented makes use of CIBERSORT [Newman et al. 2015]; in particular, we incorporate CIBERSORT analysis into additional rounds of feature selection, in order to adaptively select probe sets that best characterize the disease and influence patient conditions, thanks to prediction analysis for microarrays. This classification method is used with the aim of discovering COO of prognostic subgroups, in an attempt to improve prognostication in those subgroups with wide variation in outcome and analyze the correlation between COO and patients health.

The remainder of the chapter is structured as follows: Section 2 reports the main methods used, including frameworks used to set up the experiments; Section 3 defines the experimental activities we carried out, while Section 4 discusses the results. Eventually, Section 5 presents our conclusions and perspectives.

Part of the work presented in this chapter has been published in [Bruno et al. 2018a].

2 Proposed Approach

In the context of DNA microarrays, classifying and predicting the diagnostic category of a sample based on its gene expression profile constitute a challenge, as there is a large number of inputs (genes) from which to predict classes along with a relatively small number of samples. Hence, the identification of which genes contribute towards the classification is an important task.

The goal of this study is to provide a new approach for identifying a subset of genes that influence survival rate of patients having Diffuse Large B-cell Lymphoma. The proposed approach consists of three steps: *(i)* use CIBERSORT [Newman et al. 2015] to estimate the excess of certain member cell types in a mixed-cell population and subdivide the patients in different groups w.r.t. their own cell types value; *(ii)* apply Kaplan-Meier analysis to the groups, in order to estimate the survival function from lifetime data and measure the fraction of patients living for a certain amount of time after treatment; *(iii)* identify the best separating genes for the mixture that influence the survival rate of each subgroup. In particular, in order to perform an accurate prognosis prediction, we evaluate performance of three different classification algorithms: PAM, ClaNC and Proportional Overlapping Score (POS).

2.1 Subgroup definition using CIBERSORT

CIBERSORT [Newman et al. 2015] is a method for characterizing cell heterogeneity from nearly any tissue by using their gene expression profiles. It uses a ML approach called ν-Support Vector Regression and performs a deconvolution of mixtures, useful to analyze the composition of each sample in term of percentage of tumor and noise. The output of CIBERSORT is a new estimated mixtures that expresses in percentage the relationship between genes and cell lines and, then, the composition of each genes. By using CIBERSORT, we divided patients into two groups w.r.t. the median value computed on the B-cell proportions among all patients. In particular, let P_i be a patient, $X(P_i)$ the B-cell proportion value of that patient and M the median value, we define the *"High"* group s.t. $P_i \in High \leftrightarrow X(P_i) <= M$ and the *"Low"* group s.t. $P_i \in Low \leftrightarrow X(P_i) > M$. Basically, the *High* group contains patients which B-cell proportion is greater or equal than the median, while the *Low* group contains patients which B-cell proportion is lower than the median. The resulting groups

represent a starting point of our analysis. Indeed, for each group the Kaplan-Meier analysis is computed to obtain the overall survival of the patients.

2.2 Survival Analysis

Kaplan-Meier [Altman 1990] is a method used to measure the fraction of subjects living for a certain amount of time after treatment. The Kaplan-Meier survival function is defined as the probability of surviving in a given period of time while considering time in many small intervals. Let d_i be the number of death patients at time t_i, and let n_i be the number of patients "at risk", i.e. alive patients or not censored just before t_i (a patient is censored when information is missing); the survival function *S(t)* is computed by means of the formula:

$$S(t) = \sum_{t_i \leq t} [1 - \frac{d_i}{n_i}]$$

For instance, the probability of a patient surviving two days after a chemotherapy treatment for non-Hodgkin lymphoma is computed by conditional probability [Goel et al. 2010] as following:

$$P(t) = P(t_1 t_2 | t_1)$$

where t_1 is the probability of surviving the one day and t_2 is the probability surviving the second day.

2.3 Gene Classification

In order to identify subsets of probe sets that best characterize each class with a reasonably small cross-validation error, we consider the PAM, ClaNC and POS classification techniques, widely used in problems related to cancer-gene expression studies [Mahmoud et al. 2014]. These probe sets were removed in order to analyze the variation of survival rate by comparing the survival curves.

- PAM is a statistical technique for class prediction from gene expression data using Nearest Shrunken Centroids (NSC). It is a simple, accurate and fast classifier often used to select genes directly linked with breast cancer [Orsborne et al. 2011].

- ClaNC is a classification algorithm based on NSC. It can be represented by the centroid components and pooled by standard deviation of the active genes, which are most frequent genes for each class, demonstrated to be successful in selecting genes that discriminate between multiple clinical or biological classes [Dabney 2005].

- POS is a method based on the analysis of the overlapping regions, for each gene, yielded by the intersection between gene expression intervals of different classes with the aim to denotes gene with higher discriminative power for the considered classification problem. It is able to achieve interesting results in genes selection to increase the diagnostic value of gene expression data for colorectal cancer [Mahmoud et al. 2014].

PAM and ClaNC are based on NSC, which, in turn, is one of the most frequently used classification methods for high-dimensional data, such as microarray data [B. Y. Choi et al. 2017]. NSC selects "good" genes according to two factors: *within class distance* and *between classes distance*. When expression levels of a gene for all samples in the same class are fairly consistent with a small variance, but are largely different among samples of different classes, the gene is considered a good candidate for classification because it has discriminant information for different classes. Genes whose expression levels do not significantly differ between the classes will have their centroids reduced to the overall centroids, effectively removing them from the classification procedure [Klassen et al. 2009].

3 Experimental Setting

In the following we illustrate our experimental activities; in particular, we describe the dataset of use and the evaluation criteria adopted.

3.1 Dataset description

We conduct our experiments on the public available dataset taken from the Gene Expression Omnibus (GEO)[1], a database consisting of microarray, next generation sequencing (NGS) and other high-throughput data. In particular, we tested our method on GSE23501 dataset composed of DNA methylation signatures define molecular subtypes of Diffuse Large B-cell Lymphoma (DLBCL), characterized by probe

[1] https://www.ncbi.nlm.nih.gov/geo/

sets represented on GeneChip Human Genome U133 Plus 2.0 Array. The screening population consisted of 69 DLBCL cases in the 16–92 age range; for each patient, age, overall survival, molecular subtype, gender and treatment are known. In particular, each patient received the same treatment (R-CHOP) and this guarantees the correctness of our analysis. We used the LM22, signature matrix designed by Newman *et al* [Newman et al. 2015]. LM22 contains 547 genes that distinguish 22 human hematopoietic cell phenotypes including seven T-cell types naive and memory B-cells. These cells are highly relevant since they can kill tumor cells, or in some cases promote them growth. Precisely, we focus on B-cells memory (a type of lymphocytes) that are part of the adaptive immune system, a specific defense [Shaffer et al. 2002]. In order to perform a foreign comparison according to [Newman et al. 2015], we converted probes of references matrix (LM22) to HUGO gene symbols.

3.2 Evaluation

We used *log-rank test* and *F-test* for comparing the survival distributions of two samples. The log-rank test is based on the null hypothesis that there is no difference regarding survival among two distributions. In log-rank test we calculated the expected number of events in each group, i.e. E_1 and E_2 while O_1 and O_2 are the total number of observed events in each group, respectively. The statistic test is:

$$p = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Log-rank tests were computed within a level of significance of 5% [Bland et al. 2004].

Log-rank test may be invalid or less significant if the survival curves cross because of an increased probability of type II error [Bland et al. 2004].
It has been proved that, among ten analyses with crossed survival curves, eight are non-statistically significant and two are statistically significant. This study shows that five studies had overlapping survival curves that might be explainable for insignificant findings of the interventions [Chai-Adisaksopha et al. 2016]. For this reason, especially for the comparison of same curves belong to different distribution, we included F-test [Kao et al. 2008] in our analysis, a statistics tool for data analysis programmed to determine whether two independent estimates of variance can be assumed to be estimates of the same variance; this allows us to perform comparison between two treatments. Let $\bar{Y}_i$ the sample mean in the i_{th} group, n_i the number

TABLE 7.1: Genes distinguishing best between High and Low classes, according to PAM analysis

ID	HIGH-SCORE	LOW-SCORE
207928_S_AT	0.0995	-0.1056
1554141_S_AT	0.0774	-0.0821
230877_AT	-0.0764	0.081
1560997_AT	-0.0651	0.0691
211821_X_AT	-0.0605	0.0642
234458_AT	-0.0581	0.0616
210607_AT	-0.0569	0.0604
240791_AT	0.0542	-0.0574
236582_AT	0.0516	-0.0547
215290_AT	-0.0473	0.0501

of observations in the i_{th} group, $\bar{Y}$ the overall mean of the data, K the number of groups, N the overall sample size; then, the formula for the F-test statistic is:

$$F = \sum_{i=1}^{K} \frac{n_i(\bar{Y}_i - \bar{Y})^2}{K-1} / \sum_{i=1}^{K}\sum_{j=1}^{n} \frac{(\bar{Y}_i - \bar{Y_i})^2}{N-K}$$

In the F-test, a value of 1 denotes the equality among variances, i.e., two independent estimates of variance belong to the same variance.

4 Discussion

Patients are divided into two groups, and for each group the Kaplan-Meier analysis is computed in order to obtain the overall survival. Comparing performances of three classification algorithms, subsets of probe sets that best characterize each class are identified and removed in order to analyze the variation of survival rate. The first classification algorithm used is PAM. A grid search is used to estimate best score value, called *threshold*, that minimize classification errors. For each value, the classifier was trained using 10-fold cross validation. Precision, Recall, Accuracy and F-measure, derived from confusion matrix, and the overall MSE (i.e., the average of the squares of the difference between the estimated centroid and observed value) were used to assess the quality of the algorithm. Table 7.1 reports a subset of 10 probe sets founded by PAM with an overall cross validation error of 45.5% [25 out of 38 *High* samples were correctly predicted (63%), while 18 out of 33 *Low* were misclassified (54%)].

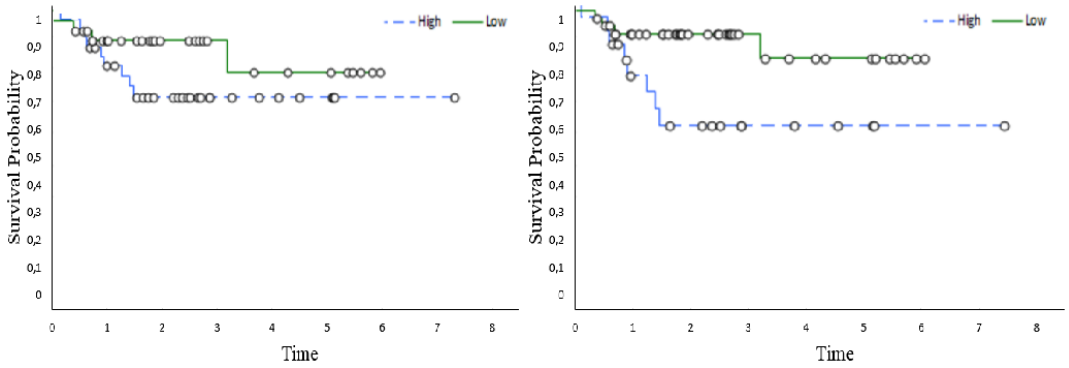


FIGURE 7.1: Plots of Kaplan-Meier product limit estimates of survival of a group of patients (on the left), and after removing genes according the PAM analysis (on the right).

These probe sets are removed from the original dataset and Kaplan-Meier analysis is performed on the resulting new subgroups with the aim of discovering relevant correlations between genes and survival rate. Notice that this operation produces changes in the composition of all cell lines and, then in the entire gene expression of each patient. For this reason, after removing probe sets, some patients were moved from one group to the other, causing a variation of survival time and probability.

Figure 7.1 shows a survival graph before and after removing these probe sets according to the PAM analysis, on the left and on the right, respectively. On the Y and X axes the estimated survival probability and the time of observation [Indrayan et al. 2001] are reported, respectively. The survival curve is drawn as a step function: the proportion surviving remains unchanged between the events, even if there are some intermediate censored observations.

Table 7.2 reports the survival time for each group in which the dataset was subdivided before (on the left) and after removing probe sets according to the PAM analysis (on the right). Rows represent the number of patients belonging to the two groups, while columns represent the number of patients observed for each group, survival time and survival probability, respectively. The number of patients in the *High* class decreases as well as the average survival rate and avarage survival probability (i.e. 60% of survival probability), w.r.t. values obtained from original dataset (on the left) (i.e. 70% of survival probability).

In order to compare the distribution of the two obtained curves (Figure 7.1), we calculated and compared the p-value according to the log-rank. Table 7.3 shows the comparison between log-rank test results obtained from original dataset (I) and

TABLE 7.2: Kaplan-Meier analysis' results before (on the left) and after removing probe sets according to PAM (on the right) in terms of average survival probability and survival time for each group (± standard deviation)

	OBSERVED	TIME	SURVIVAL
HIGH	24	3.55 (±0.47)	60%
LOW	44	5.12 (±0.33)	90%

OBSERVED	TIME	SURVIVAL
35	3.87 (±0.36)	70%
33	4.89 (±0.40)	90%

TABLE 7.3: Log-rank test computed before (I) and after the PAM analysis (II)

LOG-RANK	OBSERVED	CRITICAL VALUE	P-VALUE
I	1.93	3.84	0.17
II	4.28	3.84	0.04

the dataset after probe sets removed according to the PAM analysis (II). In particular, analysis (II) indicates a significant difference between the population survival curves (p-value 0.0391); analysis (I), instead, does not show a significant difference between the two curves (p-value 0.1650).

Average and standard deviation were used to confirm the difference between the survival curves. As reported in Table 7.4, the average of survival probability and survival time decrease from analysis (I) to analysis (II). Box-plots in Figure 7.2, computed on survival time (on the left) and on survival probability (on the right), respectively, support our previous findings: analysis (II) indicates a significant difference between the population survival curves w.r.t analysis (I) in terms of survival probability.

TABLE 7.4: Average and standard deviation computed before (I) and after PAM analysis (II) on survival probability – Y axis (right) and on survival time – X axis (left)

	AVERAGE (±*stdev*)
(I)	2.28 (±1.66)
(II)	2.21 (±1.83)

	AVERAGE (±*stdev*)
(I)	0.78 (±0.08)
(II)	0.73 (±0.13)

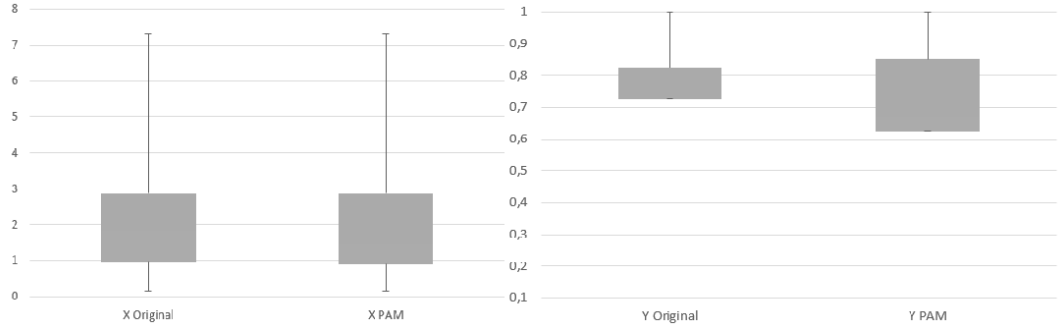


FIGURE 7.2: Box-plot computed before (I) and after PAM analysis (II) on survival probability – Y axis (right) and on survival time – X axis (left)

TABLE 7.5: Comparison the PAM, the claNC and the POS error rate (MSE) for each class

	CLASS HIGH	CLASS LOW
PAM	0.37	0.54
CLANC	0.10	0.14
POS	0.30	0.44

We performed the same procedure with the other two classification algorithms. Performance of PAM, ClaNC and POS are compared to each other, as illustrated in Table 7.5. In particular, we can observe that the overall MSE, when classifying according to ClaNC, tends to be substantially lower than the PAM error rate, in contrast for POS error.

According to each method, 10 more relevant probe sets for each class are found (see Table 7.6). Note that algorithms do not find the same probe sets.

TABLE 7.6: Comparison among the PAM, the claNC and the POS top probe sets

CLANC	PAM	POS
231192_AT	207928_S_AT	213524_S_AT
243188_AT	1554141_S_AT	201904_S_AT
227573_S_AT	230877_AT	1563203_AT
207928_S_AT	1560997_AT	241355_AT
219833_S_AT	211821_X_AT	1555801_S_AT
1554141_S_AT	234458_AT	236347_AT
221558_S_AT	210607_AT	230352_AT
215000_S_AT	240791_AT	239435_X_AT
1563127_AT	236582_AT	240529_AT
234458_AT	215290_AT	1552569_A_AT

TABLE 7.7: Kaplan-Meier analysis' results after removing probe sets according to claNC in terms of average survival time (± standard deviation)on the left and according to POS in terms of average survival time (± standard deviation) on the right

	OBSERVED	TIME	SURVIVAL
HIGH	21	3.80 (±0.47)	65%
LOW	47	4.78 (±0.32)	88%

OBSERVED	TIME	SURVIVAL
37	3.87 (±0.36)	75%
31	4.89 (±0.40)	90%

TABLE 7.8: Log-rank test computed between analysis (I), (II) and (III)

LOG-RANK	OBSERVED	CRITICAL VALUE	P-VALUE
I	4.28	3.84	0.04
II	1.18	3.84	0.28
III	2.05	3.84	0.12

Each subset of probe sets was removed from original dataset in order to perform Kaplan-Meier analysis and search for a correlation between these probe sets and survival rate of patients. Results of Kaplan-Meier analysis do not show a relevant changes after removing each probe set selected by ClaNC and POS, as indicated in Table 7.7. Indeed, the survival probability is similar to the value obtained after removing probe sets according to the PAM analysis (Table 7.2).

Our analysis is focussed only on *High* curve, that has shown a relevant change according to the PAM analysis. Table 7.8 reports the comparison between log-rank test results from resulting dataset according to the PAM analysis (I), to the ClaNC analysis (II) and to the POS analysis (III).

In particular, analysis (II) and (III) do not show a significant difference between the two curves (p-value 0.2810 and 0.1201, respectively).

As shown in Table 7.9, the average value of survival probability and survival time increases from analysis (I) to analysis (II) and analysis (III). This suggests analysis (II) and (III) do not find relevant differences between the population survival curves.

Such result is also evident from Table 7.10, which reports the result of F-test computed by comparing the *High* curve between original dataset and the resulting

TABLE 7.9: Average and standard deviation computed after the PAM analysis (I), after the ClaNC analysis (II), after the POS analysis (III) on survival probability – Y axis (up) and on survival time – X axis (down)

	AVERAGE ($\pm stdev$)
(I)	0.73 ($\pm$0.13)
(II)	0.76 ($\pm$0.12)
(III)	0.79 ($\pm$0.09)

	AVERAGE ($\pm stdev$)
(I)	2.21 ($\pm$1.83)
(II)	2.43 ($\pm$1.80)
(III)	2.26 ($\pm$1.65)

TABLE 7.10: F-test results according X (Survival time) and Y (Survival probability) axes

	F-TEST		
	PAM (I)	CLANC (II)	POS (III)
X	0.83	0.85	0.84
Y	0.42	0.51	0.46

dataset, according to each classification methods used. The F-test value increases from analysis (I) (F-test = 0.42) to analysis (II) (F-test = 0.51) and analysis (III) (F-test = 0.46), especially for survival probability axis (Y).

Our findings suggest that PAM achieves the best result with lower values of F-test, implying that the distributions of the two curves (before and after analysis (I)) are not equal.

5 Conclusion

We developed a risk model for class prediction from high-dimensional gene expression data. In particular, we first used CIBERSORT to estimate the excess of certain member cell types in a mixed-cell population, and subdivided the patients in different groups with respect to their own cell type value. In a second phase, we performed Kaplan-Meier survival analysis in order to understand which group has more chances to survive after the same treatment. We employed different statistical techniques for class prediction from gene expression data in order to detect a set of Cells-of-Origin of disease for each prognostic subgroup. Results obtained are affected by different probe set proportion between signature matrix and mixture. Indeed, only four probe sets over ten found according the PAM analysis is present in

LM22, only one according the ClaNC analysis and no probe set according the POS analysis.

As far as future works are concerned, a new signature matrix that includes more probe sets could improve our results, and better define the correlation between genes and survival rate of patients.

PART IV

Towards more Explainable AIs: Analyze the Neural Network Decision-making Process.

1 Introduction to Explainable AIs

Interpreting the decision-making processes of neural networks can be of great help at enhancing the diagnostic capabilities and providing direct patient- and process-specific support to diagnosis and disease classification. Indeed, ML and DL-based methods can be more difficult to explain and justify to human users when compared to classic analytical models [Shaban-Nejad et al. 2021]. Indeed, although a great predictive ability, most of these techniques output complex information networks, which makes the decision process difficult to explain (i.e., the well-known "black box" problem) [Anguita-Ruiz et al. 2020].

Recently, several approaches were proposed to understand the behavior of ML and DL-based systems, referred to as XAI. This is an important tool in medicine and healthcare to ensure that the results obtained by AI methods are sound, correct, and justifiable in order to help healthcare providers in making better decisions [Shaban-Nejad et al. 2021]. Thus, providing an explainability in the process of making qualified decisions, can improve and justify the treatment and health intervention, and help in translating inferred knowledge into particular hypotheses that can be tested with real-life experiments.

In the present part, we describe two approaches based on the gradient visualization techniques performed to analyze the CNNs-based model and to identify mechanisms and motivations steering the neural networks decisions in classification task. In particular, in Chapter 8 and Chapter 9 we will illustrate a XAI approaches to provide an explainability of multiple-disease classification from heatmaps or hot-spot maps and Chest X-ray images, respectively.

2 Explainable AIs: Related work

Among several approaches, the attention mechanism is one of the most widely used in DL research in the last decade, which is a method used to analyze the model's capability and highlight the most relevant information used in the prediction task. Park et al. [S. Park et al. 2018] proposed a novel disease prediction method named EHR History-based prediction using Attention Network. The approach was based on the RNN to predict vascular disease from EHR data. They included a GradCAM to visualize the class specific attention-weights and to provide interpretable results

by explicitly weighting significant features and visualizing them. Similarly, Meng et al. [Meng et al. 2019] proposed the use of GradCAM to highlight lesion regions on retinal images as a means of assisting ophthalmologists in making diagnoses. These images were fed in the CNN to improve the classification performance. However, these methods present some limitations; indeed, the generated heatmaps were basically qualitative, and not informative enough to specify which concepts have been detected. An improvement was provided using semantically explanation from visual representation [B. Zhou et al. 2018] to decompose the evidence for a prediction for image classification into semantically interpretable components, each with an identified purpose, a heatmap, and a ranked contribution. Hu et al. [H. Hu et al. 2019] proposed an attention-based DL framework, named DeepHINT, to provide mechanistic explanations on accurate prediction of HIV performed on the genomic sequence data. This approach was used to reveal important sequence positions from prediction results and thus provide important insights about the observed genomic. Choi et al. [E. Choi et al. 2016] introduced a reverse time attention model (RETAIN), which uses two attention models to detect significant clinical variables within past visits (e.g. key diagnoses). RETAIN is able to preserve interpretability, mimic the behavior of healthcare providers, and incorporate sequential information.

Other approaches are focused on analyzing the level of contribution of the input features to the output prediction to provide interpretability to ML models. Shrikumar et al. [Shrikumar et al. 2017] proposed a novel approach named Deep Learning Important FeaTures, which is a backpropagation-based interpretability approach. In particular, the approach calculates the output prediction of a network on a specific input by backpropagation algorithm to discover the importance of each feature. Lundberg et al. [Lundberg et al. 2017] proposed SHapley Additive exPlanations (SHAP) framework which uses a combination of feature contributions and game theory for breaking down the prediction with the aim to show the impact of each input feature. The overall rating of the feature contribution to the model is then achieved by aggregating the SHAP values over observations.

CHAPTER 8

Understanding Automatic Diagnosis and Classification Processes with Data Visualization

Contents

1 Introduction

In this chapter, we propose a novel method based on heatmaps and hot-spot maps for analyzing the internal processes performed by a network during the training for a classification task; the aim is to identify the most important elements that will influence the network's decisions. In particular, we rely on gradient visualization techniques to produce a coarse localization map highlighting the image regions most likely to be referred to when taking the classification decision; the identified areas are then removed from the images in order to check whether the classification performance changes.

We evaluate the proposed approach over different "instances", such as gene expression or other clinical data. Gene expression represents the amount of RNA produced in a cell under different biological states and clinical data provide both health-related information associated with patient care and features related to disease that are relevant and helpful to diagnosis classification, and even, early prediction process. Both instances consist of a set of "attributes" characterizing each patient. The attributes can be defined as: (1) proportion of genes in each gene expression and (2) clinical information related to treatment, pathology and patient characteristics.

The herein proposed method is based on a hybrid approach that relies on data visualization techniques and visual explanations for classification via a model based on CNNs; this lead to more transparent and explainable result process achievements.

The remainder of the chapter is structured as follows. In Section 2 we provide a detailed description of our approach, that has been assessed via a careful experimental activity reported in Chapter 6; we analyze and discuss results in Section 3, eventually drawing our conclusions in Section 4.

Part of the work presented in this chapter has been published in [Bruno, Pierangela et al. 2020].

2 Proposed Approach

The herein proposed approach is illustrated in Figure 8.1:

1. Data Reduction using PCA, in order to reduce the dataset dimension;
2. Data Visualization for each patient, through:
    - a heatmap to visualize proportion level of each attribute;

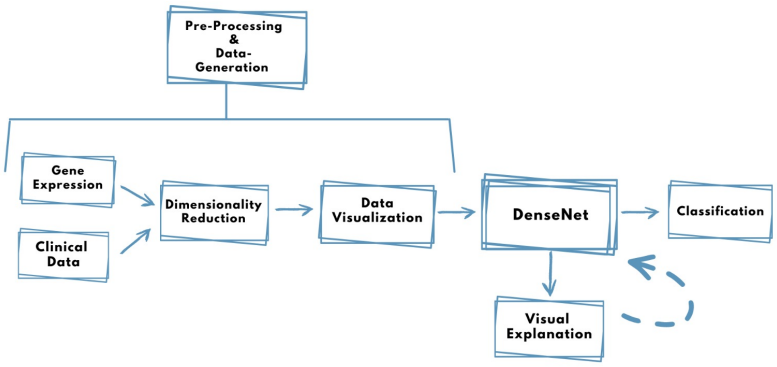


FIGURE 8.1: Workflow of the proposed framework. Dimensionality reduction is performed on Gene Expression or Clinical Data. The reduced dataset is transformed into images and diagnosis classification is performed using DenseNet 169. GradCAM and Guided GradCAM are generated during CNN training. In brackets the manuscript section.

- a hot-spot map, based on Getis-Ord Gi*, to visualize spatial information level of each attribute;

3. Disease Classification using CNNs;

4. Visual Explanations using GradCAM [Selvaraju et al. 2017] to indicate the discriminative image regions used by the CNN;

5. Verification of patients clusterization by survival analysis.

A detailed description of data processing and manipulation as well as classification steps and training phase is reported in Chapter 6, Section 4 and Section 5, respectively.

2.1 Visual Explanations

We used GradCAM [Selvaraju et al. 2017] to identify visual features in the input able to explain result process achieves during the classification. The overall structure of GradCAM is showed in Figure 8.2. In particular, it uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron. GradCAM is applied to a trained neural network with fixed weights. Given a class of interest c, let y^c the raw output of the neural network, that is, the value obtained before the application of softmax used to trasform the raw score into a probability. GradCAM performs the following three steps:

1. **Compute Gradient** of y_c w.r.t. feature maps activation A^k, for any arbitrary k, of a convolutional layer (i.e., $\frac{\delta y^c}{\delta A^k}$). This gradient value depends on the input

image chosen; indeed, the input image determines both the feature maps A^k and the final class score y^c that is produced.

2. **Calculate Alphas by Averaging Gradients** over the width dimension (indexed by i) and the height dimension (indexed by j) to obtain neuron importance weights α_k^c, as follows:

$$\alpha_k^c = \frac{1}{Z} \overbrace{\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\delta y^c}{\delta A_{i,j}^k}}_{\text{gradients via backprop}},$$

where Z is a constant (i.e., number of pixels in the activation map).

3. **Calculate Final GradCAM Heatmap** by performing a weighted combination of the feature map activations A^k as follows:

$$L_{GradCAM}^c = ReLU \underbrace{(\sum_k \alpha_k^c A^k)}_{\text{linear combination}},$$

where α_k^c is a different weight for each k, and $ReLU$ operation used to emphasize only the positive values and to convert the negative values into 0.

We also computed Guided GradCAM as proposed in [Selvaraju et al. 2017] which uses Guided Backpropagation [Springenberg et al. 2014]. It visualizes gradients w.r.t. the image where negative gradients are suppressed when backpropagating through ReLU layers (see Figure 8.2).

2.2 Survival Analysis

According to the classification results, we grouped patients of all datasets into two diagnostic groups. For each group we computed the Kaplan-Meier [Altman 1990] analysis in order to obtain the overall survival using:

1. Images generated on original dataset;
2. Images generated after removing the highlighted regions identified by Grad-CAM and Guided-GradCAM.

We compared the survival curves computed on 1 and 2 images to show if there is a statistical difference after removing the highlighted regions and then if the selected elements can be used as the diagnosis markers. The probability of survival is computed as:

$$P(t) = P(s_1 s_2 | s_1)$$

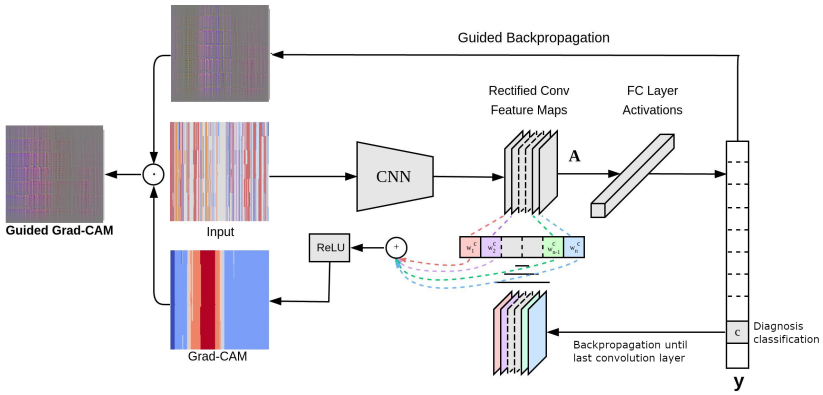


FIGURE 8.2: An example of GradCAM structure inspired by [Selvaraju et al. 2017]. Given an image, and a category ("Diagnosis c") as input, it forward propagates the image through the model to obtain the raw class scores before softmax. The gradients are set to zero for all classes except the desired class (Diagnosis c), which is set to 1. This signal is then backpropagated to the rectified convolutional feature map (A) of interest, where it allows to compute the coarse Grad-CAM localization (blue heatmap). Finally, the heatmap is multiplied with guided backpropagation to get Guided Grad CAM visualization which are both high-resolution and class-discriminative [Selvaraju et al. 2017].

where s_1 is the probability of surviving after the first day and s_2 is the probability of surviving after the second day.

2.3 Tests

Figure 8.3 lists the different approaches that we performed during our experiments. For all the approaches we used both heatmap and hot-spot map as input of CNN after the application of data reduction and data visualization.

- *Test A*: Apply GradCAM to diagnosis on both heatmap and hot-spot map using DenseNet 169.
- *Test B*: Perform diagnosis on a reduced dataset generated by removing the elements corresponding to the highlighted ares identified by Test A.
- *Test C*: Apply Guided-GradCAM to the classification performed by DenseNet, identify a set of the elements corresponding to the highlighted ares and remove from the original image input. Re-test on reduced dataset.
- *Test D*: Perform survival analysis and comparison among survival curves obtained from *Test A*-*Test B* and *Test C*

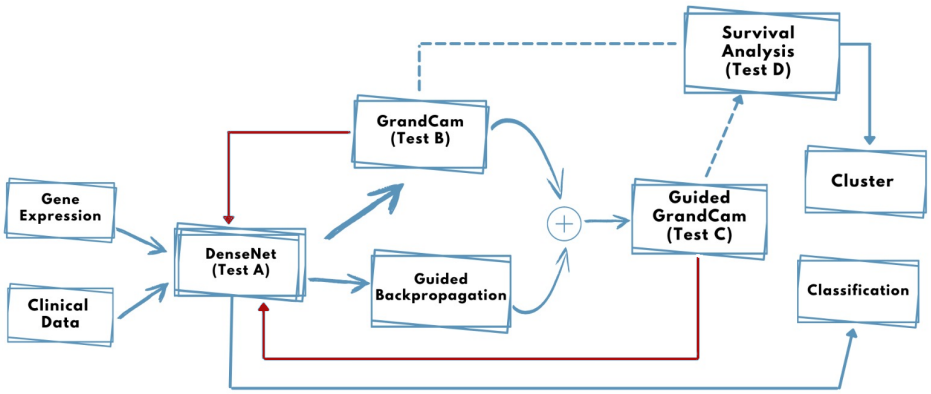


FIGURE 8.3: Workflow of the test description. Each test is performed on heatmap and hot-spot map. Red arrows indicate network re-tested using "new" dataset obtained by removing highlighted elements founded by visual explanations technique.

2.4 Performance Metrics

As reported in Chapter 6, we assessed the effectiveness of our approach using Recall, Precision anf F1-score. We also used *log-rank test* for comparing the survival distributions of two groups. The statistic test z is:

$$z = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

where E_1 and E_2 are the expected number of events in each group while O_1 and O_2 are the total number of observed events in each group. Log-rank tests were computed within a level of significance of 5% [Bland et al. 2004].

3 Results and Discussion

The experimental activities performed in Chapter 6 showed promisingly results using both gene expression and clinical data; in particular classification accuracy results higher when the visual representation is based on hotspot maps rather than heatmaps, confirming the importance of spatial information.

In Test B and Test C we selected and removed the 40% of highlighted elements; this threshold appeared to be the best compromise between data readability and quality. Figure 8.4 shows an example of the results. Both tests show a worsening of Recall value; DenseNet 169 trained on new heatmaps achieves on the 0.9454 and the 0.9354 of Recall after Test B and Test C, respectively, w.r.t. 0.9889 obtained on the

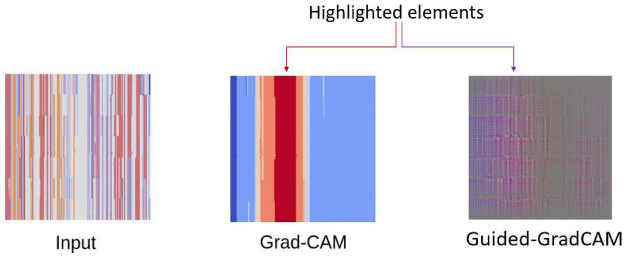


FIGURE 8.4: Example of the markers identification on a heatmap using GradCAM and Guided-GradCAM. Elements highlighted in red for GradCAM and in purple for Guided-GradCAM are considered key features in classification process.

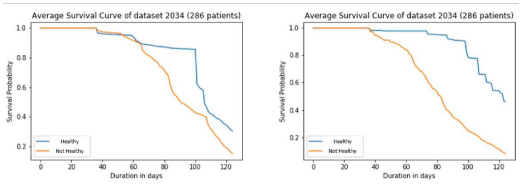


FIGURE 8.5: Plots of Kaplan-Meier product limit estimates of survival of a group of patients (on the left), and after removing genes according the Test B (on the right) computed on Breast Cancer (GSE2034) dataset.

original images. The similar worsening is shown using hot-spot maps, indeed the accuracy does not exceed the 0.96 in all the experiments. In general, a substantial decrease of Recall is shown using both Heatmap and Hot-spot map; this suggests that Test B and Test C are able to identify the important elements involved in the training process and, consequently, responsibility for this diminishment is due to images cutting by which we removed the peculiar characteristic of the disease. Figure 8.5 shows an example of survival graph computed according to the Test A and Test B, on the left and on the right, respectively.

Table 8.1 shows the comparison between log-rank test results obtained from original dataset (I), the resulting dataset after Test B (II) and after Test C (III). In particular, Test B and Test C indicate a significant difference between the population survival curves (p-value 0.0391 and p-value 0.0194); Test A, instead, does not show a significant difference between the two curves (p-value 0.1670).

4 Conclusion

We made use of GradCAM [Selvaraju et al. 2017] to analyze the internal processes performed by a neural network during the training phase, with the aim of improving explainability in the process of making qualified decisions; more in detail, we

TABLE 8.1: Log-rank test computed according Breast Cancer (GSE2034) dataset before (I) and after *Approach B (II) and after* Approach C (III) on survival probability (i.e Y axis) (right) and on survival time (i.e. X axis) (left)

LOG-RANK	OBSERVED	CRITICAL VALUE	P-VALUE
I	2.73	3.84	0.17
II	3.98	3.84	0.04
III	4.18	3.84	0.02

try to identify the most important regions that influence the network's decisions. We tested the proposed approach over DenseNet 169 trained for the task of automatic medical diagnosis based on images representing high-dimensional gene expression and clinical data. The use of images for representing data presents two relevant advantages: first of all, it significantly eases and improves data visualization; furthermore, it allows taking advantage of effective techniques explicitly geared towards image processing in order to perform classification tasks.

Starting from numerical "raw" data, we make use of PCA to reduce the dimensions, getting rid of redundant or irrelevant information and paving the way to a proper 2-D image-based representation.

Experimental results show that not only our proposal is comparable to current state-of-the-art methods, proving to be effective and robust, but it is also able to identify specific regions that are crucial in the neural network decision-making process, thus improving explainability. Indeed, classification accuracy is lower when highlighted regions are removed from the input images; this suggests the importance of these areas in disease classification and the possibility to consider the set of elements identified as potential disease markers.

In general, as one might expect, dataset quality, along with careful dimensionality reductions and feature selection schemes, are of great importance for effective CNNs, and subsequently for accurate disease predictions. Hence, the choice of the most relevant and informative feature subset for training a model is the basis of robust and competitive models.

In contexts where early and accurate medical diagnosis of specific pathologies are essential, our method proves that data visualization combined with ML techniques can be used to analyze high-dimensional multivariate data and automatically discover new bio-markers by interpreting network decisions.

As future work is concerned, we plan to investigate misclassification errors and build more robust classifiers with better generalization.

CHAPTER 9

Understanding Automatic Pneumonia Classification using Chest X-ray images

Contents

1 Introduction

The Novel Coronavirus, that reportedly started to infect human individuals at the end of 2019, rapidly caused a pandemic, as the infection can spread quickly from individual to individual in the community [Öztürk et al. 2020]. Signs of infection include respiratory symptoms, fever, cough and dyspnea. In more serious cases, the infection can cause Pneumonia, severe acute respiratory syndrome, septic shock, multi-organ failure, and death [Organization et al. 2020; McKeever 2020]. These symptoms are similar to those caused by other respiratory infections diseases, such as tuberculosis (TB). TB is a chronic lung disease caused by bacterial infection, and is one of the top-10 leading causes of death [Rahman et al. 2020]. Both TB and COVID-19 primarily attack the lungs, although with different incubation period from exposure to disease, and ill people show similar symptoms such as cough, fever and difficulty breathing [Motta et al. 2020]. In such scenarios, correct classification of the diseases is crucial for ensuring that patients get the right treatment. Early and automatic diagnoses can provide support to clinicians in such complex decisions as well as to control the epidemic, paving the way to timely referral of patients to quarantine, rapid intubation of serious cases in specialized hospitals, and monitoring of the spread of the disease. Since the disease heavily affects human lungs, analyzing Chest X-ray images of the lungs may prove to be a powerful tool for disease investigation. Several methods have been proposed in the literature in order to perform disease classification from Chest X-ray images, especially based on DL approaches [Zotin et al. 2019; H. Liu et al. 2019; Bullock et al. 2019]. Notably, in this context, solutions featuring interpretability and explainability approaches can significantly help at improving disease classification and providing context-aware assistance and understanding. Indeed, interpreting the decision-making processes of neural networks can be of great help at enhancing the diagnostic capabilities and providing direct patient- and process-specific support to diagnosis. However, interpretability and explainability represent critical points in approaches based on DL models, that achieved great results in disease classification.

In this work, we investigate the use of CNNs with the aim to perform multiple-disease classification from Chest X-ray images. Diseases that are a matter of concern for our experiments are COVID-19 and TB Pneumonia. Notably, although these diseases are characterized by pulmonary inflammation caused by different pathogens,

TB Pneumonia has similar clinical symptoms to COVID-19 [Motta et al. 2020] that could affect a proper diagnosis and treatment plan. Moreover, we include in our experiments Healthy patients to learn how they differ from symptomatic patients.

We analyze the CNNs-based model to identify the mechanisms and the motivations steering neural networks decisions in classification task. In particular, we use gradient visualization techniques to produce coarse localization maps highlighting the image regions most likely to be referred to by the model when the classification decision is taken. The highlighted areas are then used to discover (i) patterns in Chest X-ray images related to a specific disease, and (ii) correlation between these areas and classification accuracy, by analyzing a possible performance worsening after their removal.

The remainder of the chapter is structured as follows. We first briefly report on related work in Section 2; in Section 3 we then provide a detailed description of our approach, that has been assessed via a careful experimental activity, which is discussed in Section 4; we analyze and discuss results in Section 5, eventually drawing our conclusions in Section 6.

Part of the work proposed in this chapter has been published in [Bruno et al. 2020a].

2 Related work in disease classification

In this section we present state-of-the-art methods used to perform disease classification through Chest X-ray images. DL-based models recently achieved promising results in image-based disease classification. These models, such as CNNs [Moccia et al. 2019a; Colleoni et al. 2019; Spadea et al. 2019; Zaffino et al. 2019], are proven to be appropriate and effective when compared to conventional methods; indeed, CNNs currently represent the most widely used method for image processing. Abbas et al. [Abbas et al. 2020] proposed a DL approach (DeTraC) to perform disease classification using X-ray images. The approach was used to distinguish COVID-19 X-ray images from normal ones, achieving an accuracy of 95.12%. An improvement in terms of binary classification accuracy was presented by Ozturk et al. [Ozturk et al. 2020]. The authors proposed a DL model (DarkCovidNet) for automatic diagnosis of COVID-19 based on Chest X-ray images. They both performed a binary and multi-class classification, dealing with patients with COVID-19, no-findings and

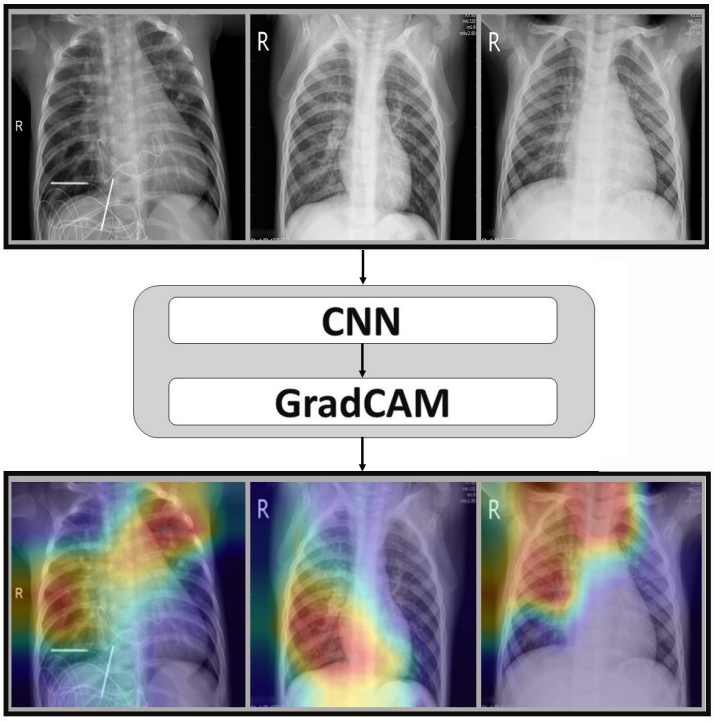


FIGURE 9.1: Workflow of the proposed framework. Chest X-ray images are used to train the CNN. The last convolution layer of the CNN is used as input of the GradCAM approach to provide the corresponding visual explanations (i.e., the regions of input that are "important" for classification).

Pneumonia. The accuracy achieved is of 98.08% and 87.02%, respectively. Similarly, Wang et al. [Linda Wang et al. 2020] proposed a DL-based approach (COVID-Net) to detect distinctive abnormalities in Chest X-ray images among patients with non-COVID-19 viral infections, bacterial infections, and healthy patients, achieving an overall accuracy of 92.6%. All the approaches showed limitations related to low number of image samples and imprecise localization on the chest region. More accurate localization of model's prediction was proposed by Mangal et al. [Mangal et al. 2020] and Haghanifar et al. [Haghanifar et al. 2020]. The authors proposed a DL-based approach to classify COVID-19 patients from others/normal ones. They also generated saliency maps to show the classification score obtained during the prediction and to validate the results.

In this work, we propose the use of DL approach to perform multiple disease classification using Chest X-ray. Additionally, we take advantage of a novel technique for analyzing the internal processes and the decision performed by a neural network during the training phase.

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201
Conv	112 × 112	7 × 7 conv, stride 2		
Pooling	56 × 56	3 × 3 max pool, stride 2		
DB (1)	56 × 56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
TL (1)	56 × 56	1 × 1 conv		
	28 × 28	2 × 2 average pool, sride 2		
DB (2)	28 × 28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
TL (2)	28 × 28	1 × 1 conv		
	14 × 14	2 × 2 average pool, sride 2		
DB (3)	14 × 14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$
TL (3)	14 × 14	1 × 1 conv		
	7 × 7	2 × 2 average pool, sride 2		
DB (4)	7 × 7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$
CL	4 × 1	7 × 7 global average pool		
		4D fully-connected, softmax		

TABLE 9.1: Architecture of the networks DenseNet-121, DenseNet-169 and DenseNet-201. More in detail, Conv stands for convolution, DB for Dense Block, TL for Transition Layer, CL for Classification Layer.

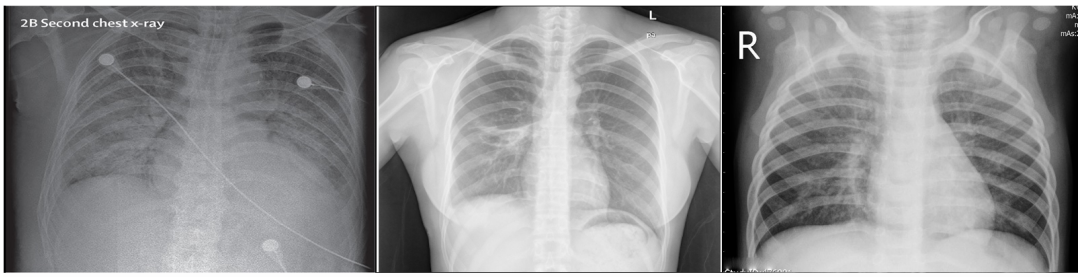
3 Proposed Approach

In the following we illustrate the herein proposed approach: we first describe how the classification methods are designed and then how the choices are explained by properly highlighting regions that are considered discriminant for the classification. A detailed description of visual explanation technique is reported in the Section 2.1 of Chapter 8.

3.1 Classification

Given that symptoms of COVID-19 pneumonia can be similar to those caused by other respiratory illnesses, including TB, distinguishing between them is extremely important, especially during a pandemic. Therefore, our purpose aims at providing methods for automatically identifying the “correct" condition of a given patient based on her Chest X-ray images, and also some details about the reasons for the resulting classifications.

The herein proposed approach, illustrated in Figure 9.1, is based on: (i) Multiple-disease classification using CNNs (trained according to 2 similar-based symptoms diseases, namely COVID-19 and TB Pneumonia), and (ii) Visual Explanations using



(a) COVID-19 (b) TB Pneumonia (c) Healthy patient

FIGURE 9.2: Example of frontal-view Chest X-ray images for the treated pathologies.

GradCAM [Selvaraju et al. 2017] to indicate the discriminative image regions used by the CNN.

To classify patients, we used and compared the results of three neural networks chosen on the basis of the good performance obtained on the *ImageNet* data set over several competitions [Rosebrock 2017]. In particular, we make use of DenseNet 121, DenseNet 169 and DenseNet 201.

DenseNet networks [G. Huang et al. 2017] are made of dense blocks, as shown in Table 9.1, where for each layer the inputs are the feature maps of all the previous layers with the aim to improve the information flow on 224×224 input images. More in detail, for convolutional layers with kernel size 3×3, each side of the inputs is zero-padded by one pixel to keep the feature-map size fixed. The layers between two contiguous dense blocks are referred as transition layers for convolution and pooling, which contain 1×1 convolution and 2×2 average pooling. A 1×1 convolution is introduced as a bottleneck layer before each 3×3 convolution to reduce the number of input feature-maps, and thus to improve computational efficiency. At the end of the last dense block, a global average pooling and a softmax classifier are applied.

4 Experimental Protocol

We describe next the setting of the experimental analysis performed in order to assess the viability of our approach.

4.1 Dataset description

For the experimental analysis we used datasets provided by Cohen et al. [Cohen et al. 2020], Kermany et al. [Kermany et al. 2018] for COVID-19 and normal patients, respectively, and Jaeger et al. [Jaeger et al. 2013], Candemir et al. [Candemir et al. 2013] and Wang et al. [Xiaosong Wang et al. 2017] for patients which suffer from TB. The datasets consist of several X-ray extracted from various online publications and websites. Examples of X-ray images are shown in Figure 9.2.

In particular, we considered only 3 specific categories distributed as follows:

1. COVID-19 Pneumonia, counting 434 patients
2. Tuberculosis Pneumonia, counting 336 patients
3. Healthy patients, counting 1667 patients

In order to obtain a valid classification and avoid majority class selection, we properly made use of data augmentation techniques to over-sample imbalanced data an obtain an equal number of samples in abundant class. More specifically, we performed:

- Translating medical images: shift the region of interest with respect to the center of the training images;
- Rotating medical images: rotate the training images by a random amount of degrees;
- Flipping medical images: use randomized flipping, through which the image information is mirrored horizontally or vertically.

4.2 Training phase

The dataset was split into training (80%) and testing (20%) sets; the 20% of the training set is used as validation set, in order to monitor the training process and prevent overfitting. We started relying on the work on [Mangal et al. 2020]. We used CheXNet model [H. Wang et al. 2018] as pre-trained weights to improve the robustness of our approach and sigmoid classifier, as suggested by [Mangal et al. 2020].

All experiments have been performed on a machine equipped with a GeForce GTX 970 GPU.

Fine-tuning For the training phase we performed hyperparameters optimization. DenseNets was trained with both optimizers Adam and SGD and for each

TABLE 9.2: Validation Recall for the 3 tested neural networks after 10-fold cross-validation for each dataset. Most significant results hare highlighted.

DATASET	DenseNet 121	DenseNet 169	DenseNet 201
COVID-19	0.94	**0.95**	0.90
TB Pneumonia	0.79	**0.84**	0.66
Healthy patients	**0.94**	0.88	0.87

optimizer 4 learning rate were tried. The best performance is obtained with the following configuration, trained for 200 epochs: SGD optimizer, learning rate 10^{-4}, batch size 4, and binary cross-entropy as loss function.

The configuration of networks was modified in terms of the number of nodes or levels to optimize the performance. We empirically changed the number of layers and we trimmed network size by pruning nodes to improve computational performance and identify those nodes which would not noticeably affect network performance. However, since we performed the experiments using well-know networks already optimized, we achieved the best performance using the standard configuration as originally proposed by respective authors.

We performed 10-fold cross-validation in order to choose the parameter value that gives the lowest cross-validation average error; experiments were performed on the very same machine with the same configuration of the other approaches.

4.3 Performance Metrics

We assessed the effectiveness of our approach by measuring AUC and Recall, especially focusing on the last one; indeed, in this context, the most important thing is to minimize False Negatives (i.e., disease is present but is not identified).

The AUC, which is 1 for a perfect system [Marıén et al. 2010], is computed according to the Equation 4.1 and Equation 4.2 described in Chapter 4, Section 3. Closer a curve approaches the top left corner, then better is the performance of the system.

Recall (see formula in Chapter 6, Section 5.4) considers prediction accuracy among only actual positives and explain how correct our prediction is among all people.

TABLE 9.3: AUC values for the 3 tested neural networks after 10-fold cross-validation for each dataset. Most significant results hare highlighted.

DATASET	DenseNet 121	DenseNet 169	DenseNet 201
COVID-19	0.96	**0.97**	0.92
TB Pneumonia	**0.95**	0.92	0.94
Healthy patients	0.95	**0.96**	0.93

5 Results and Discussion

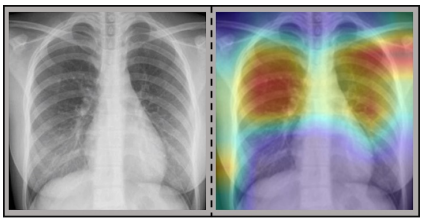
In the following, we first discuss the quality of the classifications performed by our models, and then assess the visual explanations of the choices, as provided by Grad-CAM.

5.1 Classification Performance

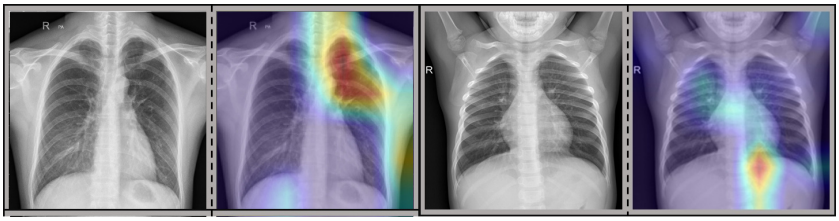
Table 9.2 and Table 9.3 report classification results after 10-fold cross-validation for all datasets in terms of Recall and AUC, respectively. Each network is evaluated on a test set generated performing random data augmentation. Even though similar results are achieved in all DenseNet-based experiments, DenseNet 169 shows one of the most efficient architecture: if reports AUC mean value of 0.95 and Recall mean value of 0.89 over all the classes: hence, it was the one selected for the study.

The herein proposed approach achieves promisingly results: the best performance was achieved by DenseNet 169 on COVID-19 dataset (i.e., Recall mean value: 0.95 and AUC: 0.97), outperforming results on Healthy patients (i.e., Recall mean value of 0.88 and AUC: 0.96), and especially results in classifying TB Pneumonia (i.e., Recall mean value: 0.84 and AUC: 0.92).

A more thorough analysis shows that that TB Pneumonia is often confused with COVID-19; this is not surprising, given the overlapping imaging characteristics that leads to a Recall mean value below 0.85 in all experiments performed. Furthermore, it is worth noting that, in general, the extraction of CT scan images from published articles, rather than from actual sources, might lessen image quality, thus affecting performance of the ML model.



(A) COVID-19



(B) TB Pneumonia (C) Healthy patients

FIGURE 9.3: Visual example of achieved results. For each diagnostic class, we show raw Chest X-ray image (left) and GradCAM result (right). Images on the right sides highlight the most important areas involved in the classification process.

5.2 Assessing explanations from GradCAM

As already mentioned, we make use of GradCAM for highlighting the most significant regions w.r.t. classification. A visual inspection of the GradCAM output confirms the quality of the model; indeed, it exhibits strong classification criteria in the Chest region (see Figure 9.3). In particular, red areas refer to the parts where the attention is strong, while blue areas refer to weaker attention. In general, the warmer the color, the more important to the network are the highlighted features.

In order to confirm that the portions identified by GradCAM are actually significant, we performed both quantitative and qualitative analyses.

As for a quantitative assessment, we selected and removed, for each dataset, the 40% of highlighted elements as suggested in [Bruno et al. 2020b]. A substantial decrease of Recall (on average around 5%) is shown using COVID-19, TB Pneumonia and Healthy patients (i.e., p-value < 0.05 for paired t-test computed before and after images cutting). This result suggests that GradCAM is actually able to identify the important elements involved in the training process and, consequently, responsibility for this diminishment is due to images cutting that removed some peculiar characteristic of the disease.

Furthermore, we took advantage from the TB dataset, that feature labels assigned by expert clinicians to the areas of the images they considered relevant for the classification. The current clinical practice consists in visually inspecting and

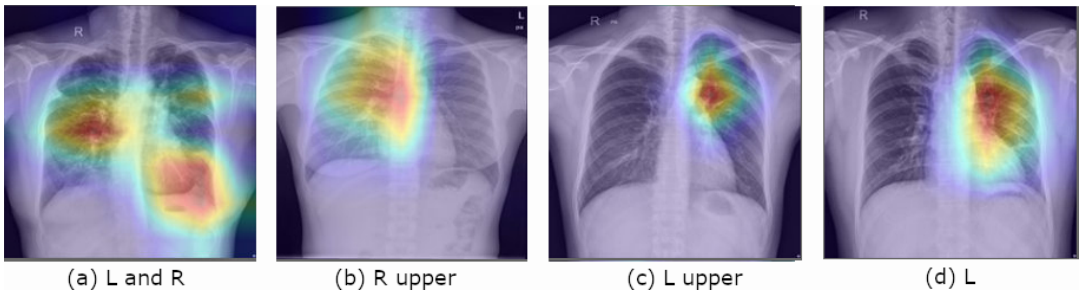


FIGURE 9.4: Example of results obtained after GradCAM application. For each image, label assigned by clinicians is provided.

evaluating Chest X-ray to identify abnormality of the lung and define the clinical condition of the patient. To the best of our knowledge no clinical evaluations are provided for COVID-19 patients: this is why this analysis is limited to TB patients only. We wanted to assess the capability of GradCAM in identifying relevant elements involved in the training process and discovering potentially bio-markers able to suggests clinical condition, and also check to what extent the features identified by GradCAM overlap with those used by the clinicians.

Clinicians provided the evaluation of Chest X-ray of TB patients identifying the approximate location of the abnormality in lung; in particular, clinical categories are subdivided into individual lung (left, right) or bilateral lungs and, for each of these three categories, clinicians specified areas involved in TB disease (i.e., upper, lower, middle lung). As an example, Figure 9.3b shows an image of the lungs of a patient suffering from pulmonary TB: clinicians assigned the label “L upper" starting from the image on the left, while the image on the right reports the highlighted region by GradCAM. A visual example of the preliminary results is shown in Figure 9.4: they show that what GradCAM highlights in our approach based on DenseNet-169 coincides with at least 60% of the area suggested by clinicians, thus identifying a great number of suggested areas. However, the accuracy does not reach a considerably high value due to (i) the capability of our network in classifying TB patients, as discussed in Section 5.1, (ii) the way areas identified by clinicians are labelled, as labels are not defined at pixels level nor delimited, but rather by means of texts, thus resulting in non-precise delimitation. Furthermore, result suggest that there might be some cases in which the artificial networks and the human operators just focus on different features; this can be subject of future investigations.

6 Conclusion

In this work we exploit the use of CNNs and visual explanation techniques to estimate diagnosis using Chest X-ray and to analyze the internal processes performed by a neural network during the training phase with the aim of improving explainability in the process of making qualified decisions. Basically, we try to identify the most important regions that influence the network's decisions.

We fine-tuned the approach by means of accurate experimental activities; in particular, we classified three different datasets (i.e., two for ill and one for healthy patients), and three different CNNs for the classification.

Experimental results show that our proposal is robust and it is able to identify specific regions that are crucial in the neural network decision-making process, thus improving explainability. Indeed, classification accuracy is lower when highlighted regions are removed from the input images; this suggests the importance of these areas in disease classification and the possibility to consider the set of elements identified as potential disease markers. We also discussed the relationships between the highlighted visual features suggested by GradCAM and the abnormalities identified by clinicians on TB Chest X-ray.

In context where early and accurate medical diagnosis of specific pathologies are essential, our method proves that visual explanation method combined with ML techniques can be used to provide solid disease classifications and automatically discover new bio-markers by interpreting network decisions.

As future work is concerned, we aim to investigate misclassification errors and improve the generalization capability of the model. Our efforts will also focus on including clinical evaluations of COVID-19 Chest X-ray; we also plan to find medical expertise at pixels or coordinates level to judge and better assess the quality of the regions highlighted by explanation approach. With this respect, we also plan to explore explanations methods other than GradCAM.

PART V

Conclusions and Perspectives

CHAPTER 10

Conclusion

In healthcare, early and accurate diagnosis of specific disease is crucial in ensuring the most effective treatments and the most appropriate therapeutic decisions. Clinicians' diagnostic practices involve reviewing medical histories of patients and analysing clinical data and medical imaging; unfortunately, this requires a lot of significant efforts.

In this context, AI-based approaches opened up new perspectives in assisting clinicians in the diagnosis of diseases and treatment decisions, revealing to be capable of extracting important pieces of information and identifying latent patterns in high-dimensional data that are of great help at solving complex relevant problems in different domains.

We aimed at contributing in the field, by proposing a ML-based approach to classify patients in prognostic sub-groups from gene expression data and to predict a risk model. Since clinical and omics data such as gene expression are characterized by high dimensionality, the analysis and manipulation of these data can be difficult. To overcome these limitations, we proposed a combination of data reduction, data visualization and DL-based classification; more in detail, we reduced the dimension of the huge amount of data collected from patients and we converted them into heatmap capable to visualize proportion level of each attribute with the goal to rely on CNN-based approaches for diagnosis classification.

In the last decades, DL approaches have also been used to analyze medical images, showing excellent performance in several applications, such as segmentation and object detection (e.g., surgical tools). We presented a DL-based method to perform vascular segmentation of ilio-femoral district, showing the robustness of our approach to images jeopardized by the presence of catheters, noise and other artifacts. Furthermore, we proposed a DL-based approach to segment multi-instances of instruments in laparoscopic procedures. Our workflow showed that DL in combination with temporal information and instrument likelihood prediction can improve the generalizability and robustness of multi-instance segmentation.

Eventually, to better support clinicians in the actual clinical practice and to validate the results provided by AI-based approach, we worked on explainability approaches, making a step towards a satisfactory interpretation of the defined models. Specifically, we relied on specific attention mechanism to identify visual features in the input which can potentially explain result process achieves during the classification task.

CHAPTER 11

Perspective

In this study we explored a variety of AI-based techniques to support automatic medical diagnosis and investigate pathological mechanisms. Thanks to these techniques – such as ML and DL – it is possible to find insights into diagnostics, care processes and treatment variability, identifying disease-related biomarkers and improving patient outcome.

The herein proposed approaches represent a starting point for more general and complex ways towards precision medicine and health. An interesting perspective might be provided by the use medical images jointly to patients' records (e.g., clinical and omics data) in the decision-making process, simulating the diagnostic workflow of a clinician and potentially improving both robustness and interpretability of the automated processes.

Interpretability and explainability, such as understanding the most important elements for a model at hand, have gained particular attention in the scientific community, as they are crucial for increasing the trust in predictions and recommendations made by AI-based systems. Explainable AI (XAI) techniques arise with this specific aim: provide better and more interpretable results. However, most of the current approaches just provide localization maps, for instance highlighting image features that are relevant for the prediction model, and fail in converting them in human-interpretable suggestions of use for an "acceptable" explanation of the automated decision processes. In this context, interesting perspective can be provided

by including rule-based approaches able to convert such highlighted areas, making them "readily usable" in several application domain.

Another promising direction of work is to enhance disease prediction and prevention by using digital representations of patients (i.e., the use of *Digital Twins* in healthcare) which could be essential for tasks like defining customized treatment decisions, predicting outcomes, identifying the best drug set among all possibilities for a certain disease, virtually testing several treatments (in advance), thus reducing impact on patients (such as toxicity) and risks. Digital twins supported by AI-based approach could significantly help clinicians to properly optimize the performance of treatment plans, based on specific patient characteristics, allowing to anticipate potential complications.

Thanks to their ability to collect and analyze data in a clear and concise way, reducing imprecise judgments, AI-based approaches could drive progress in medicine, supporting clinicians in disease diagnosis and treatment.

In the future, these approaches can be used to significantly increase automation in the identification of the most accurate diagnosis and best therapeutic strategies; assistance to healthcare providers such as radiologists at detecting cancer from medical images will likely improve dramatically, meaning that the way such specialists and clinicians work will naturally change even faster than in the past years.

Bibliography

Abadi, Martıén et al. (2016). "Tensorflow: A system for large-scale machine learning". In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283.

Abbas, Asmaa, Mohammed M Abdelsamea, and Mohamed Medhat Gaber (2020). "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network". In: *arXiv preprint arXiv:2003.13815*.

Abul-Husn, Noura S and Eimear E Kenny (2019). "Personalized medicine and the power of electronic health records". In: *Cell* 177.1, pp. 58–69.

Aggarwal, Charu C et al. (2018). "Neural networks and deep learning". In: *Springer* 10, pp. 978–3.

Ahmed, Zeeshan (2020). "Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis". In: *Human Genomics* 14.1, pp. 1–5.

Alashwal, Hany et al. (2019). "The application of unsupervised clustering methods to Alzheimer's Disease". In: *Frontiers in computational neuroscience* 13, p. 31.

Aličković, Emina and Abdulhamit Subasi (2017). "Breast cancer diagnosis using GA feature selection and Rotation Forest". In: *Neural Computing and Applications* 28.4, pp. 753–763.

Allan, Max et al. (2015). "Image based surgical instrument pose estimation with multi-class labelling and optical flow". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 331–338.

Allan, Max et al. (2020). "2018 Robotic Scene Segmentation Challenge". In: *arXiv preprint arXiv:2001.11190*.

Almugren, Nada and Hala Alshamlan (2019). "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification". In: *IEEE Access* 7, pp. 78533–78548.

Alonso-Caneiro, David et al. (2018). "Use of convolutional neural networks for the automatic segmentation of total retinal and choroidal thickness in OCT images". In:

Altman, Douglas G (1990). *Practical statistics for medical research*. CRC press.

Ando, Tatsuya et al. (2002). "Fuzzy Neural Network Applied to Gene Expression Profiling for Predicting the Prognosis of Diffuse Large B-cell Lymphoma". In: *Japanese Journal of Cancer Research* 93.11, pp. 1207–1212.

Anguita-Ruiz, Augusto et al. (2020). "eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research". In: *PLOS Computational Biology* 16.4, e1007792.

Awad, Mariette and Rahul Khanna (2015). "Support vector regression". In: *Efficient learning machines*. Springer, pp. 67–80.

Azimian, Hamidreza, Rajni V Patel, and Michael D Naish (2010). "On constrained manipulation in robotics-assisted minimally invasive surgery". In: *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics*. IEEE, pp. 650–655.

Barto, Andrew G (1997). "Reinforcement learning". In: *Neural systems for control*. Elsevier, pp. 7–30.

Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool (2006). "Surf: Speeded up robust features". In: *European conference on computer vision*. Springer, pp. 404–417.

Becht, Etienne et al. (2019). "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1, p. 38.

Bengio, Yoshua, Ian Goodfellow, and Aaron Courville (2017). *Deep learning*. Vol. 1. MIT press Massachusetts, USA:

Bergel, Alexandre (2020). "The Artificial Neuron". In: *Agile Artificial Intelligence in Pharo*. Springer, pp. 37–51.

Bernal, Jorge et al. (2015). "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized Medical Imaging and Graphics* 43, pp. 99–111.

Bernal, Jorge et al. (2017). "Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge". In: *IEEE transactions on medical imaging* 36.6, pp. 1231–1249.

Bharti, Kusum Kumari and Pramod Kumar Singh (2014). "A three-stage unsupervised dimension reduction method for text clustering". In: *Journal of Computational Science* 5.2, pp. 156–169.

Biscione, Valerio and Jeffrey Bowers (2020). "Learning Translation Invariance in CNNs". In: *arXiv preprint arXiv:2011.11757*.

Bland, J Martin and Douglas G Altman (2004). "The logrank test". In: *Bmj* 328.7447, p. 1073.

Bobak, Carly A, Alexander J Titus, and Jane E Hill (2019). "Comparison of common machine learning models for classification of tuberculosis using transcriptional biomarkers from integrated datasets". In: *Applied Soft Computing* 74, pp. 264–273.

Bodenstedt, Sebastian et al. (2018). "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery". In: *arXiv preprint arXiv:1805.02475*.

Bogdanova, Rositsa, Pierre Boulanger, and Bin Zheng (2016). "Depth perception of surgeons in minimally invasive surgery". In: *Surgical innovation* 23.5, pp. 515–524.

Bradley, Andrew P (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7, pp. 1145–1159.

Brazma, Alvis et al. (2001). "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data". In: *Nature genetics* 29.4, pp. 365–371.

Brisimi, Theodora S et al. (2018). "Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach". In: *Proceedings of the IEEE* 106.4, pp. 690–707.

Bruno, Pierangela and Francesco Calimeri (2019). "Using Heatmaps for Deep Learning based Disease Classification". In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp. 1–7.

— (2020a). "Understanding Automatic Pneumonia Classification Using Chest X-Ray Images". In: pp. 37–50.

Bruno, Pierangela, Francesco Calimeri, and Aldo Marzullo (2018a). "Classification and survival prediction in Diffuse Large B-cell Lymphoma by gene expression

profiling". In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer, pp. 166–178.

Bruno, Pierangela, Marte Cinzia, and Francesco Calimeri (2020b). "Understanding Automatic COVID-19 Classification using Chest X-ray images". In: *Proceedings of the Italian Workshop on Explainable Artificial Intelligence co-located with 19th International Conference of the Italian Association for Artificial Intelligence*.

Bruno, Pierangela et al. (2018b). "Using cnns for designing and implementing an automatic vascular segmentation method of biomedical images". In: *International Conference of the Italian Association for Artificial Intelligence*. Springer, pp. 60–70.

Bruno, Pierangela et al. (2020c). "Data Reduction and Data Visualization for Automatic Diagnosis using Gene Expression and Clinical Data". In: *Artificial Intelligence in Medicine*, p. 101884.

Bruno, Pierangela et al. (2020). "Understanding Automatic Diagnosis and Classification Processes with Data Visualization". In: pp. 1–6.

Bullock, Joseph, Carolina Cuesta-Lázaro, and Arnau Quera-Bofarull (2019). "XNet: A convolutional neural network (CNN) implementation for medical X-Ray image segmentation suitable for small datasets". In: *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Vol. 10953. International Society for Optics and Photonics, 109531Z.

Calimeri, Francesco et al. (2019). "A Logic-Based Framework Leveraging Neural Networks for Studying the Evolution of Neurological Disorders". In: *Theory and Practice of Logic Programming*, pp. 1–45.

Candemir, Sema et al. (2013). "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration". In: *IEEE transactions on medical imaging* 33.2, pp. 577–590.

Canny, JF (1987). "A computational approach to edge detection, Readings in computer vision: issues, problems, principles, and paradigms". In: *IEEE Trans. Pattern*.

Chai-Adisaksopha, Chatree et al. (2016). "A systematic review of using and reporting survival analyses in acute lymphoblastic leukemia literature". In: *BMC hematology* 16.1, p. 17.

Che, Zhengping et al. (2018). "Recurrent neural networks for multivariate time series with missing values". In: *Scientific reports* 8.1, pp. 1–12.

Chen, Runpu et al. (2020). "Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data". In: *Bioinformatics* 36.5, pp. 1476–1483.

Chen, Yifei et al. (2016). "Gene expression inference with deep learning". In: *Bioinformatics* 32.12, pp. 1832–1839.

Chervitz, Stephen A et al. (2011). "Data standards for Omics data: the basis of data sharing and reuse". In: *Bioinformatics for Omics Data*. Springer, pp. 31–69.

Choi, Byeong Yeob, Eric Bair, and Jae Won Lee (2017). "Nearest shrunken centroids via alternative genewise shrinkages". In: *PloS one* 12.2, e0171068.

Choi, Edward et al. (2016). "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism". In: *Advances in Neural Information Processing Systems*, pp. 3504–3512.

Christ, Patrick (2017). "LiTS Liver Tumor Segmentation Challenge (LiTS17)". In: *URL https://competitions. codalab. org/competitions/17094*.

Clark, David J and Hui Zhang (2020). "Proteomic approaches for characterizing renal cell carcinoma". In: *Clinical Proteomics* 17.1, pp. 1–18.

Cleveland, William S and Susan J Devlin (1988). "Locally weighted regression: an approach to regression analysis by local fitting". In: *Journal of the American statistical association* 83.403, pp. 596–610.

Codella, Noel CF et al. (2018). "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)*. IEEE, pp. 168–172.

Cohen, Joseph Paul et al. (2020). "Covid-19 image data collection: Prospective predictions are the future". In: *arXiv preprint arXiv:2006.11988*.

Colleoni, Emanuele et al. (2019). "Deep Learning Based Robotic Tool Detection and Articulation Estimation With Spatio-Temporal Layers". In: *IEEE Robotics and Automation Letters* 4.3, pp. 2714–2721.

Corey, Kristin M et al. (2018). "Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study". In: *PLoS medicine* 15.11, e1002701.

Crum, William R, Oscar Camara, and Derek LG Hill (2006). "Generalized overlap measures for evaluation and validation in medical image analysis". In: *IEEE transactions on medical imaging* 25.11, pp. 1451–1461.

Dabney, Alan R (2005). "Classification of microarrays to nearest centroids". In: *Bioinformatics* 21.22, pp. 4148–4154.

Dahl, George E, Tara N Sainath, and Geoffrey E Hinton (2013). "Improving deep neural networks for LVCSR using rectified linear units and dropout". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 8609–8613.

Dakua, Sarada Prasad et al. (2019). "Moving object tracking in clinical scenarios: application to cardiac surgery and cerebral aneurysm clipping". In: *International journal of computer assisted radiology and surgery* 14.12, pp. 2165–2176.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Dice, Lee R (1945). "Measures of the amount of ecologic association between species". In: *Ecology* 26.3, pp. 297–302.

Dosovitskiy, A et al. (2019). "& Brox, T.(2015). Flownet: Learning optical flow with convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2758–2766.

Dosovitskiy, Alexey et al. (2015). "Flownet: Learning optical flow with convolutional networks". In: *IEEE International Conference on Computer Vision*, pp. 2758–2766.

Drozdzal, Michal et al. (2016). "The importance of skip connections in biomedical image segmentation". In: *Deep Learning and Data Labeling for Medical Applications*. Springer, pp. 179–187.

Du, Xiaofei et al. (2018). "Articulated multi-instrument 2-D pose estimation using fully convolutional networks". In: *IEEE transactions on medical imaging* 37.5, pp. 1276–1287.

Elsayed, Nelly, Anthony S Maida, and Magdy Bayoumi (2018). "Deep Gated Recurrent and Convolutional Network Hybrid Model for Univariate Time Series Classification". In: *arXiv preprint arXiv:1812.07683*.

Esteva, Andre et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639, pp. 115–118.

Evans, Jae A (July 1999). *Electronic medical records system*. US Patent 5,924,074.

Fenster, Araon and Bernard Chiu (2006). "Evaluation of segmentation algorithms for medical imaging". In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, pp. 7186–7189.

Ford, Elizabeth et al. (2020). "Can the Use of Bayesian Analysis Methods Correct for Incompleteness in Electronic Health Records Diagnosis Data? Development of a Novel Method Using Simulated and Real-Life Clinical Data". In: *Frontiers in Public Health* 8, p. 54.

Fraz, Muhammad Moazam et al. (2012). "An ensemble classification-based approach applied to retinal blood vessel segmentation". In: *IEEE Transactions on Biomedical Engineering* 59.9, pp. 2538–2548.

Fritzsche, K et al. (2003). "Automated model based segmentation, tracing and analysis of retinal vasculature from digital fundus images". In: *State-of-The-Art Angiography, Applications and Plaque Imaging Using MR, CT, Ultrasound and X-rays* 29, pp. 225–298.

Fu, Mei R et al. (2020). "Precision health: A nursing perspective". In: *International journal of nursing sciences* 7.1, pp. 5–12.

Funke, Isabel et al. (2019). "Video-based surgical skill assessment using 3D convolutional neural networks". In: *International journal of computer assisted radiology and surgery* 14.7, pp. 1217–1225.

Gadosey, Pius Kwao et al. (2020). "Sd-unet: Stripping down u-net for segmentation of biomedical images on platforms with low computational budgets". In: *Diagnostics* 10.2, p. 110.

Gajula, MNVP (2012). "Its Time to Integrate Multi Omics Data to understand Real Biology". In: *International Journal of Systems, Algorithms & Applications* 2, pp. 31–34.

Garcelon, Nicolas et al. (2020). "Electronic health records for the diagnosis of rare diseases". In: *Kidney International* 97.4, pp. 676–686.

Garcıéa-Peraza-Herrera, Luis C et al. (2016). "Real-time segmentation of non-rigid surgical tools based on deep learning and tracking". In: *International Workshop on Computer-Assisted and Robotic Endoscopy*. Springer, pp. 84–95.

Goel, Manish Kumar, Pardeep Khanna, and Jugal Kishore (2010). "Understanding survival analysis: Kaplan-Meier estimate". In: *International journal of Ayurveda research* 1.4, p. 274.

Gold, Steven, Anand Rangarajan, et al. (1996). "Softmax to softassign: Neural network algorithms for combinatorial optimization". In: *Journal of Artificial Neural Networks* 2.4, pp. 381–399.

González, Cristina, Laura Bravo-Sánchez, and Pablo Arbelaez (2020). "ISINet: An Instance-Based Approach for Surgical Instrument Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 595–605.

Goyal, Manu (2019). "Natural Data-augmentation for Skin Lesions (ISIC-2017 Challenge Dataset)". In: *Mendeley Data, V1*.

Guo, Changlu et al. (2020). "SA-UNet: Spatial Attention U-Net for Retinal Vessel Segmentation". In: *arXiv preprint arXiv:2004.03696*.

Guo, LinJie et al. (2020). "Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos)". In: *Gastrointestinal endoscopy* 91.1, pp. 41–51.

Gupta, Mehak et al. (2019). "Obesity Prediction with EHR Data: A deep learning approach with interpretable elements". In: *arXiv*, arXiv–1912.

Hafiz, Abdul Mueed and Ghulam Mohiuddin Bhat (2020). "A survey on instance segmentation: state of the art". In: *International Journal of Multimedia Information Retrieval*, pp. 1–19.

Haghanifar, Arman, Mahdiyar Molahasani Majdabadi, and Seokbum Ko (2020). "Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning". In: *arXiv preprint arXiv:2006.13807*.

Hannun, Awni Y et al. (2019). "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network". In: *Nature medicine* 25.1, p. 65.

Hawkes, Peter W (2004). *Advances in imaging and electron physics*. Elsevier.

He, Juncai et al. (2018). "ReLU deep neural networks and linear finite elements". In: *arXiv preprint arXiv:1807.03973*.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Kaiming et al. (2017). "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

Health, National Institutes of et al. (2007). "Understanding emerging and re-emerging infectious diseases: biological sciences curriculum study NIH Curriculum Supplement Series National Institutes of Health". In: *Bethesda, MD*.

Hedström, Gustaf et al. (2015). "The impact of age on survival of diffuse large B-cell lymphoma–a population-based study". In: *Acta Oncologica* 54.6, pp. 916–923.

Hira, Zena M and Duncan F Gillies (2015). "A review of feature selection and feature extraction methods applied on microarray data". In: *Advances in bioinformatics* 2015.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Hoefsloot, HCJ et al. (2020). "Multiset data analysis: ANOVA simultaneous component analysis and related methods". In: *Comprehensive Chemometrics [Recurso electrónico]: chemical and biochemical data analysis. Volume 1*. Elsevier, pp. 465–478.

Höhlein, Kevin et al. (2020). "A comparative study of convolutional neural network models for wind field downscaling". In: *Meteorological Applications* 27.6, e1961.

Hoover, AD, Valentina Kouznetsova, and Michael Goldbaum (2000). "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response". In: *IEEE Transactions on Medical imaging* 19.3, pp. 203–210.

Howard, Andrew G et al. (2017). "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861*.

Howard, Brian E, Beate Sick, and Steffen Heber (2009). "Unsupervised assessment of microarray data quality using a Gaussian mixture model". In: *BMC bioinformatics* 10.1, p. 191.

Howe, Jessica L et al. (2018). "Electronic health record usability issues and potential contribution to patient harm". In: *Jama* 319.12, pp. 1276–1278.

Hsu, H. and A. L. Peter (2007). "Paired t test". In: *Wiley encyclopedia of clinical trials*, pp. 1–3.

Hu, Hailin et al. (2019). "DeepHINT: understanding HIV-1 integration via deep learning with attention". In: *Bioinformatics* 35.10, pp. 1660–1667.

Hu, Kai et al. (2018). "Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function". In: *Neurocomputing* 309, pp. 179–191.

Huang, Gao et al. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huang, Huimin et al. (2020). "Unet 3+: A full-scale connected unet for medical image segmentation". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1055–1059.

Hundman, Kyle et al. (2018). "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, pp. 387–395.

Iacobucci, Ilaria et al. (2019). "Genomic subtyping and therapeutic targeting of acute erythroleukemia". In: *Nature genetics* 51.4, pp. 694–704.

Ilg, Eddy et al. (2017). "FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks". In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470. (Visited on 12/16/2019).

Indrayan, A and SB Surmukaddam (2001). "Measurement of community health and survival analysis". In: *Medical Biostatistics* 7, pp. 232–42.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167*.

Isensee, Fabian and Klaus H Maier-Hein (2020). "OR-UNet: an Optimized Robust Residual U-Net for Instrument Segmentation in Endoscopic Images". In: *arXiv preprint arXiv:2004.12668*.

Jaeger, Stefan et al. (2013). "Automatic tuberculosis screening using chest radiographs". In: *IEEE transactions on medical imaging* 33.2, pp. 233–245.

Jakobsen, Janus Christian et al. (2017). "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts". In: *BMC medical research methodology* 17.1, p. 162.

Jha, Debesh et al. (2020). "Doubleu-net: A deep convolutional neural network for medical image segmentation". In: *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, pp. 558–564.

Jin, Yueming et al. (2019). "Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 440–448.

Kahl, G (2015). "Minimum information about a high-throughput nucleotide sequencing experiment (MINSEQE)". In: *The Dictionary of Genomics, Transcriptomics and Proteomics. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA*.

Kalyani, Chenigaram, Kama Ramudu, and Ganta Raghotham Reddy (2020). "Enhancement and Segmentation of Medical Images Using AGCWD and ORACM." In: *International Journal of Online & Biomedical Engineering* 16.13.

Kao, Lillian S and Charles E Green (2008). "Analysis of variance: is there a difference in means and what does it mean?" In: *Journal of Surgical Research* 144.1, pp. 158–170.

Karim, Fazle et al. (2017). "LSTM fully convolutional networks for time series classification". In: *IEEE Access* 6, pp. 1662–1669.

Kate, Rohit J et al. (2019). "Continual Prediction from EHR Data for Inpatient Acute Kidney Injury". In: *arXiv preprint arXiv:1902.10228*.

Kermany, Daniel, Kang Zhang, Michael Goldbaum, et al. (2018). "Labeled optical coherence tomography (OCT) and Chest X-Ray images for classification". In: *Mendeley data* 2.2.

Ketkar, Nikhil (2017). "Stochastic gradient descent". In: *Deep learning with Python*. Springer, pp. 113–132.

Khalid, Samina, Tehmina Khalil, and Shamila Nasreen (2014). "A survey of feature selection and feature extraction techniques in machine learning". In: *2014 Science and Information Conference*. IEEE, pp. 372–378.

Khan, Asifullah et al. (2019). "A survey of the recent architectures of deep convolutional neural networks". In: *Artificial Intelligence Review*, pp. 1–62.

Khoshhali, Mehri et al. (2012). "Predicting the survival time for diffuse large B-cell lymphoma using microarray data". In: *Journal of Molecular and Genetic Medicine: an International Journal of Biomedical Research* 6, p. 287.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Klassen, Myungsook and Nyunsu Kim (2009). "Nearest Shrunken Centroid as Feature Selection of Microarray Data." In: *CATA*, pp. 227–232.

Kletz, Sabrina et al. (2019). "Identifying surgical instruments in laparoscopy using deep learning instance segmentation". In: *IEEE International Conference on Content-Based Multimedia Indexing*. IEEE, pp. 1–6.

Kress, Markus (2010). *Intelligent business process optimization for the service industry*. KIT Scientific Publishing.

Krishnankutty, Binny et al. (2012). "Data management in clinical research: an overview". In: *Indian journal of pharmacology* 44.2, p. 168.

Kroese, Dirk P et al. (2013). *Cross-entropy method. Encyclopedia of Operations Research and Management Science*.

Kuhn, Harold W (1955). "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1-2, pp. 83–97.

Lazar, Cosmin et al. (2012). "A survey on filter techniques for feature selection in gene expression microarray analysis". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.4, pp. 1106–1119.

Lee, Jeong Min and Milos Hauskrecht (2020). "Multi-scale temporal memory for clinical event time-series prediction". In: *International Conference on Artificial Intelligence in Medicine*. Springer, pp. 313–324.

Lenz, Georg (2013). "Novel therapeutic targets in diffuse large B-cell lymphoma". In: *EJC Suppl* 11.2, pp. 262–3.

Li, Bo et al. (2010). "Gene expression data classification using locally linear discriminant embedding". In: *Computers in Biology and Medicine* 40.10, pp. 802–810.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14, pp. 1754–1760.

Li, Ruowang et al. (2020). "Electronic health records and polygenic risk scores for predicting disease risk". In: *Nature Reviews Genetics*, pp. 1–10.

Li, Xiaomeng et al. (2018). "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes". In: *IEEE transactions on medical imaging* 37.12, pp. 2663–2674.

Liang, Christine A et al. (2019). "Proteomics Analysis of FLT3-ITD Mutation in Acute Myeloid Leukemia Using Deep Learning Neural Network". In: *Annals of Clinical & Laboratory Science* 49.1, pp. 119–126.

Lin, Bingxiong et al. (2016). "Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey". In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 12.2, pp. 158–178.

Lin, Shan et al. (2019). "Automatic Sinus Surgery Skill Assessment Based on Instrument Segmentation and Tracking in Endoscopic Video". In: *International Workshop on Multiscale Multimodal Medical Imaging*. Springer, pp. 93–100.

Lin, Tsung-Yi et al. (2014). "Microsoft COCO: Common objects in context". In: *European Conference on Computer Vision*. Springer, pp. 740–755.

Liu, Han et al. (2019). "SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images". In: *Computerized Medical Imaging and Graphics* 75, pp. 66–73.

Liu, Jian et al. (2017). "Tumor gene expression data classification via sample expansion-based deep learning". In: *Oncotarget* 8.65, p. 109646.

Liu, Li et al. (2018). "Utility of inverse probability weighting in molecular pathological epidemiology". In: *European journal of epidemiology* 33.4, pp. 381–392.

Liu, Yuying et al. (2020). "An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery". In: *IEEE Access* 8, pp. 78193–78201.

Lowe, Rohan et al. (2017). "Transcriptomics technologies". In: *PLoS computational biology* 13.5, e1005457.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems*, pp. 4765–4774.

Madhavan, Subha et al. (2018). "Art and challenges of precision medicine: interpreting and integrating genomic data into clinical practice". In: *American Society of Clinical Oncology Educational Book* 38, pp. 546–553.

Mahmoud, Osama et al. (2014). "A feature selection method for classification within functional genomics experiments based on the proportional overlapping score". In: *BMC bioinformatics* 15.1, p. 274.

Maier-Hein, Lena et al. (2017). "Surgical data science for next-generation interventions". In: *Nature Biomedical Engineering* 1.9, p. 691.

Maier-Hein, Lena et al. (2021). "Heidelberg colorectal data set for surgical data science in the sensor operating room". In: *Scientific data* 8.1, pp. 1–11.

Malan, Linda et al. (2020). "Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns". In: *Nutrition Research* 75, pp. 67–76.

Mangal, Arpan et al. (2020). "CovidAID: COVID-19 Detection Using Chest X-Ray". In: *arXiv preprint arXiv:2004.09803*.

Marıén, Diego et al. (2010). "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features". In: *IEEE Transactions on medical imaging* 30.1, pp. 146–158.

Matas, Jiri et al. (2004). "Robust wide-baseline stereo from maximally stable extremal regions". In: *Image and vision computing* 22.10, pp. 761–767.

Mcdade, Kevin K, Uma Chandran, and Roger S Day (2015). "Improving cancer gene expression data quality through a TCGA data-driven evaluation of identifier filtering". In: *Cancer informatics* 14, CIN–S33076.

McDermott, Mary M et al. (2011). "Superficial femoral artery plaque and functional performance in peripheral arterial disease: walking and leg circulation study (WALCS III)". In: *JACC: Cardiovascular Imaging* 4.7, pp. 730–739.

McGinnis, J Michael et al. (2011). *Clinical data as the basic staple of health learning: Creating and protecting a public good: Workshop summary*. National Academies Press.

McGlinchy, Joe et al. (2019). "Application of UNet Fully Convolutional Neural Network to Impervious Surface Segmentation in Urban Environment from High Resolution Satellite Imagery". In: *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, pp. 3915–3918.

McKeever, Amy (2020). "Here's what coronavirus does to the body". In: *National Geographic*.

Meister, Simon, Junhwa Hur, and Stefan Roth (2018). "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Meng, Qier, Yohei Hashimoto, and Shin'ichi Satoh (2019). "Fundus image classification and retinal disease localization with limited supervision". In: *Asian Conference on Pattern Recognition*. Springer, pp. 469–482.

Miotto, Riccardo et al. (2016). "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". In: *Scientific reports* 6.1, pp. 1–10.

Mitchell, Tom M et al. (1997). "Machine learning. 1997". In: *Burr Ridge, IL: McGraw Hill* 45.37, pp. 870–877.

Moccia, Sara et al. (2018). "Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics". In: *Computer methods and programs in biomedicine* 158, pp. 71–91.

Moccia, Sara et al. (2019a). "Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images". In: *Magnetic Resonance Materials in Physics, Biology and Medicine* 32.2, pp. 187–195.

Moccia, Sara et al. (2019b). "Preterm infants' pose estimation with spatio-temporal features". In: *IEEE Transactions on Biomedical Engineering*.

Motta, Ilaria et al. (2020). "Tuberculosis, COVID-19 and migrants: preliminary analysis of deaths occurring in 69 patients from two cohorts". In: *Pulmonology* 26.4, pp. 233–240.

Mou, Luntian et al. (2019). "T-LSTM: A long short-term memory neural network enhanced by temporal information for traffic flow prediction". In: *Ieee Access* 7, pp. 98053–98060.

Muzio, Giulia, Leslie O'Bray, and Karsten Borgwardt (2020). "Biological network analysis with deep learning". In: *Briefings in Bioinformatics*.

Narayanan, NK et al. (2015). "Image segmentation based on multiple means using class division method". In: *2015 International Conference on Industrial Instrumentation and Control (ICIC)*. IEEE, pp. 1264–1267.

Newman, Aaron M et al. (2015). "Robust enumeration of cell subsets from tissue expression profiles". In: *Nature methods* 12.5, pp. 453–457.

Nguyen, Xuan Anh et al. (2019). "Surgical skill levels: Classification and analysis using deep neural network model and motion signals". In: *Computer methods and programs in biomedicine* 177, pp. 1–8.

Nielsen, Michael A (2015). *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA.

Njoku, Kelechi et al. (2020). "Metabolomic Biomarkers for Detection, Prognosis and Identifying Recurrence in Endometrial Cancer". In: *Metabolites* 10.8, p. 314.

Novianto, Sonny, Yukinori Suzuki, and Junji Maeda (2003). "Near optimum estimation of local fractal dimension for image segmentation". In: *Pattern Recognition Letters* 24.1-3, pp. 365–374.

Organization, World Health et al. (2020). "Health Topics. Coronavıérus". In: *Coronavirus: symptoms. World Health Organization, 2020a. Disponıével em: https://www. who. int/healthtopics/coronavirus# tab= tab_3. Acesso em* 7.

Oromendia, Ana et al. (2020). *Error-free, automated data integration of exosome cargo protein data with extensive clinical data in an ongoing, multi-omic translational research study.*

Orsborne, Christopher and Richard Byers (2011). "Impact of gene expression profiling in lymphoma diagnosis and prognosis". In: *Histopathology* 58.1, pp. 106–127.

Ozturk, Tulin et al. (2020). "Automated detection of COVID-19 cases using deep neural networks with X-ray images". In: *Computers in Biology and Medicine*, p. 103792.

Öztürk, Şaban, Umut Özkaya, and Mücahid Barstuğan (2020). "Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features". In: *International Journal of Imaging Systems and Technology*.

Pal, Mahesh and Giles M Foody (2010). "Feature selection for classification of hyperspectral data by SVM". In: *IEEE Transactions on Geoscience and Remote Sensing* 48.5, pp. 2297–2307.

Park, Seunghyun et al. (2018). "Interpretable prediction of vascular diseases from electronic health records via deep attention networks". In: *18th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2018*. Institute of Electrical and Electronics Engineers Inc., pp. 110–117.

Peng, Hanchuan, Fuhui Long, and Chris Ding (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8, pp. 1226–1238.

Prior, Barry M et al. (2004). "Time course of changes in collateral blood flow and isolated vessel size and gene expression after femoral artery occlusion in rats". In: *American Journal of Physiology-Heart and Circulatory Physiology* 287.6, H2434–H2447.

Qin, Fangbo et al. (2019). "Surgical Instrument Segmentation for Endoscopic Vision with Data Fusion of rediction and Kinematic Pose". In: *IEEE International Conference on Robotics and Automation*. IEEE, pp. 9821–9827.

Quang, Daniel and Xiaohui Xie (2016). "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences". In: *Nucleic acids research* 44.11, e107–e107.

Rahman, Tawsifur et al. (2020). "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization". In: *IEEE Access* 8, pp. 191586–191601.

Rakhlin, Alexander, Alex Davydow, and Sergey I Nikolenko (2018). "Land Cover Classification From Satellite Imagery With U-Net and Lovasz-Softmax Loss." In: *CVPR Workshops*, pp. 262–266.

Ren, Shaoqing et al. (2015). "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *arXiv preprint arXiv:1506.01497*.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Rosebrock, Adrian (2017). "Imagenet: Vggnet, resnet, inception, and xception with keras". In: *Mars*.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Roß, Tobias et al. (2020). "Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge". In: *Medical Image Analysis*, p. 101920.

Roy, Asim (2015). "A Classification Algorithm for High-dimensional Data." In: *Inns Conference on Big Data*, pp. 345–355.

Sajjad, Muhammad et al. (2019). "Multi-grade brain tumor classification using deep CNN with extensive data augmentation". In: *Journal of computational science* 30, pp. 174–182.

Sandberg, Rickard and Ingemar Ernberg (2005). "Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI)". In: *Proceedings of the National Academy of Sciences* 102.6, pp. 2052–2057.

Sandler, Mark et al. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520.

Sandlund, John T and Mike G Martin (2016). "Non-Hodgkin lymphoma across the pediatric and adolescent and young adult age spectrum". In: *Hematology 2014, the American Society of Hematology Education Program Book* 2016.1, pp. 589–597.

Sarikaya, Duygu, Jason J Corso, and Khurshid A Guru (2017). "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection". In: *IEEE transactions on medical imaging* 36.7, pp. 1542–1549.

Scott, David W et al. (2014). "Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue". In: *Blood* 123.8, pp. 1214–1217.

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Setio, Arnaud Arindra Adiyoso et al. (2017). "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge". In: *Medical image analysis* 42, pp. 1–13.

Shaban-Nejad, Arash, Martin Michalowski, and David L Buckeridge (2021). "Explainability and Interpretability: Keys to Deep Medicine". In: *Explainable AI in Healthcare and Medicine*, p. 1.

Shaffer, AL, Andreas Rosenwald, and Louis M Staudt (2002). "Lymphoid malignancies: the dark side of B-cell differentiation". In: *Nature Reviews Immunology* 2.12, pp. 920–933.

Shao, Yaping and Weidong Le (2019). "Recent advances and perspectives of metabolomics-based investigations in Parkinson's disease". In: *Molecular Neurodegeneration* 14.1, p. 3.

Sharma, Alok and Kuldip K Paliwal (2008). "Cancer classification by gradient LDA technique using microarray gene expression data". In: *Data & Knowledge Engineering* 66.2, pp. 338–347.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). "Learning important features through propagating activation differences". In: *arXiv preprint arXiv:1704.02685*.

Shvets, Alexey A et al. (2018). "Automatic instrument segmentation in robot-assisted surgery using deep learning". In: *IEEE International Conference on Machine Learning and Applications)*. IEEE, pp. 624–628.

Siddaiah-Subramanya, Manjunath, Kor Woi Tiang, and Masimba Nyandowe (2017). "A new era of minimally invasive surgery: progress and development of major technical innovations in general surgery over the last decade". In: *The Surgery Journal* 3.04, e163–e166.

Singla, Rohit et al. (2017). "Intra-operative ultrasound-based augmented reality guidance for laparoscopic surgery". In: *Healthcare technology letters* 4.5, pp. 204–209.

Soler-Botija, Carolina, Carolina Gálvez-Montón, and Antoni Bayes Genis (2019). "Epigenetic biomarkers in cardiovascular diseases". In: *Frontiers in genetics* 10, p. 950.

Spadea, Maria Francesca et al. (2019). "Deep Convolution Neural Network (DCNN) Multiplane Approach to Synthetic CT Generation From MR images—Application in Brain Proton Therapy". In: *International Journal of Radiation Oncology* Biology* Physics* 105.3, pp. 495–503.

Springenberg, Jost Tobias et al. (2014). "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806*.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Staal, Joes et al. (2004). "Ridge-based vessel segmentation in color images of the retina". In: *IEEE transactions on medical imaging* 23.4, pp. 501–509.

Stamate, Daniel et al. (2019). "A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort". In: *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 5.1, pp. 933–938.

Subramanian, Indhupriya et al. (2020). "Multi-omics data integration, interpretation, and its application". In: *Bioinformatics and biology insights* 14, p. 1177932219899051.

Sugimoto, Masahiro et al. (2012). "Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis". In: *Current bioinformatics* 7.1, pp. 96–108.

Sui, Daniel Z (2004). "Tobler's first law of geography: A big idea for a small world?" In: *Annals of the Association of American Geographers* 94.2, pp. 269–277.

Szegedy, Christian et al. (2016). "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826.

Taha, Isra N and Alexandra Naba (2019). "Exploring the extracellular matrix in health and disease using proteomics". In: *Essays in Biochemistry* 63.3, pp. 417–432.

Tarek, Sara, Reda Abd Elwahab, and Mahmoud Shoman (2017). "Gene expression based cancer classification". In: *Egyptian Informatics Journal* 18.3, pp. 151–159.

Taylor, Chris F et al. (2007). "The minimum information about a proteomics experiment (MIAPE)". In: *Nature biotechnology* 25.8, pp. 887–893.

Tebani, Abdellah et al. (2016). "Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations". In: *International journal of molecular sciences* 17.9, p. 1555.

Thankaswamy-Kosalai, Subazini, Partho Sen, and Intawat Nookaew (2017). "Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics". In: *Genomics* 109.3-4, pp. 186–191.

Thomas, Jaya, Sonia Thomas, and Lee Sael (2017). "DP-miRNA: An improved prediction of precursor microRNA using deep learning model". In: *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, pp. 96–99.

Tieleman, Tijmen and Geoffrey Hinton (2012). "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude". In: *COURSERA: Neural networks for machine learning* 4.2, pp. 26–31.

Van't Veer, Laura J et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer". In: *nature* 415.6871, pp. 530–536.

Vedula, S Swaroop and Gregory D Hager (2017). "Surgical data science: the new knowledge domain". In: *Innov Surg Sci* 2.3, pp. 109–121.

Vergara, Jorge R and Pablo A Estévez (2014). "A review of feature selection methods based on mutual information". In: *Neural computing and applications* 24.1, pp. 175–186.

Voillet, Valentin et al. (2016). "Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework". In: *BMC bioinformatics* 17.1, pp. 1–16.

Wallisch, Pascal et al. (2009). "Chapter 29 - Neural Network Part II: Supervised Learning". In: *Matlab for Neuroscientists*. Ed. by Pascal Wallisch et al. Academic Press, pp. 319–337.

Wang, Bo, Shuang Qiu, and Huiguang He (2019). "Dual encoding u-net for retinal vessel segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 84–92.

Wang, Hongyu and Yong Xia (2018). "Chestnet: A deep neural network for classification of thoracic diseases on chest radiography". In: *arXiv preprint arXiv:1807.03058*.

Wang, John (2003). "Data mining: opportunities and challenges". In:

Wang, Linda and Alexander Wong (2020). "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images". In: *arXiv preprint arXiv:2003.09871*.

Wang, Lizhi et al. (2021). "Chapter One-A Deep-forest based approach for detecting fraudulent online transaction." In: *Adv. Comput.* 120, pp. 1–38.

Wang, Rong et al. (2017). "3-D Tracking for Augmented Reality Using Combined Region and Dense Cues in Endoscopic Surgery". In: *IEEE journal of biomedical and health informatics* 22.5, pp. 1540–1551.

Wang, Sheng et al. (2017). "Accurate de novo prediction of protein contact map by ultra-deep learning model". In: *PLoS computational biology* 13.1, e1005324.

Wang, Shu-Lin, Yaping Fang, and Jianwen Fang (2013). "Diagnostic prediction of complex diseases using phase-only correlation based on virtual sample template". In: *BMC bioinformatics*. Vol. 14. S8. Springer, S11.

Wang, Shulin et al. (2006). "SVM-based tumor classification with gene expression data". In: *international conference on advanced data mining and applications*. Springer, pp. 864–870.

Wang, Xiaosong et al. (2017). "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106.

Wang, Xizhao, Yanxia Zhao, and Farhad Pourpanah (2020). *Recent advances in deep learning*.

Weichenhan, Dieter et al. (2020). "Epigenomic technologies for precision oncology". In: *Seminars in Cancer Biology*. Elsevier.

West, Mike et al. (2001). "Predicting the clinical status of human breast cancer by using gene expression profiles". In: *Proceedings of the National Academy of Sciences* 98.20, pp. 11462–11467.

Wise, Anastasia L et al. (2019). "Genomic medicine for undiagnosed diseases". In: *The Lancet* 394.10197, pp. 533–540.

Wu, Hao, Qi Liu, and Xiaodong Liu (2019). "A review on deep learning approaches to image classification and object segmentation". In: *Comput. Mater. Continua* 60.2, pp. 575–597.

Wu, Thomas D and Colin K Watanabe (2005). "GMAP: a genomic mapping and alignment program for mRNA and EST sequences". In: *Bioinformatics* 21.9, pp. 1859–1875.

Wu, Po-Yen et al. (2016). "–Omic and electronic health record big data analytics for precision medicine". In: *IEEE Transactions on Biomedical Engineering* 64.2, pp. 263–273.

Wurcel, Victoria et al. (2019). "The value of diagnostic information in personalised healthcare: A comprehensive concept to facilitate bringing this technology into healthcare systems". In: *Public health genomics* 22.1-2, pp. 8–15.

Xiao, Yawen et al. (2018). "A deep learning-based multi-model ensemble method for cancer prediction". In: *Computer methods and programs in biomedicine* 153, pp. 1–9.

Xie, Hongtao et al. (2019). "Automated pulmonary nodule detection in CT images using deep convolutional neural networks". In: *Pattern Recognition* 85, pp. 109–119.

Yamada, Ryo et al. (2020). "Interpretation of omics data analyses". In: *Journal of Human Genetics*, pp. 1–10.

Yang, Luanyi and Zeshui Xu (2019). "Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning". In: *International Journal of Machine Learning and Cybernetics* 10.3, pp. 591–601.

Yang, Siyuan et al. (2018). "Automatic coronary artery segmentation in X-ray angiograms by multiple convolutional neural networks". In: *Proceedings of the 3rd International Conference on Multimedia and Image Processing*, pp. 31–35.

Yin, Qijin et al. (2019). "DeepHistone: a deep learning approach to predicting histone modifications". In: *BMC genomics* 20.2, pp. 11–23.

Yu, Xiang-Tian and Tao Zeng (2018). "Integrative analysis of omics big data". In: *Computational Systems Biology*. Springer, pp. 109–135.

Zaffino, Paolo et al. (2019). "Fully automatic catheter segmentation in MRI with 3D convolutional neural networks: application to MRI-guided gynecologic brachytherapy". In: *Physics in Medicine & Biology* 64.16, p. 165008.

Zampieri, Guido et al. (2019). "Machine and deep learning meet genome-scale metabolic modeling". In: *PLoS computational biology* 15.7, e1007084.

Zeiler, Matthew D (2012). "Adadelta: an adaptive learning rate method". In: *arXiv preprint arXiv:1212.5701*.

Zhang, Jianpeng et al. (2019a). "Attention residual learning for skin lesion classification". In: *IEEE transactions on medical imaging* 38.9, pp. 2092–2103.

Zhang, Jianpeng et al. (2019b). "Medical image classification using synergic deep learning". In: *Medical image analysis* 54, pp. 10–19.

Zhang, Li et al. (2018). "Applying 1-norm SVM with squared loss to gene selection for cancer classification". In: *Applied Intelligence* 48.7, pp. 1878–1890.

Zhang, Zhongheng (2016). "Multiple imputation with multivariate imputation by chained equation (MICE) package". In: *Annals of translational medicine* 4.2.

Zhao, Yingdong and Richard Simon (2010). "Gene expression deconvolution in clinical samples". In: *Genome medicine* 2.12, pp. 1–3.

Zhou, Bolei et al. (2018). "Interpretable basis decomposition for visual explanation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 119–134.

Zhou, Zongwei et al. (2018). "Unet++: A nested u-net architecture for medical image segmentation". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, pp. 3–11.

Zhu, Bin et al. (2017). "Integrating clinical and multiple omics data for prognostic assessment across human cancers". In: *Scientific reports* 7.1, pp. 1–13.

Zoph, Barret et al. (2018). "Learning transferable architectures for scalable image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710.

Zotin, Aleksandr et al. (2019). "Lung boundary detection for chest X-ray images classification based on GLCM and probabilistic neural networks". In: *Procedia Computer Science* 159, pp. 1439–1448.