UNIVERSITY OF CATANIA

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

Deep Learning for data analysis on specific contexts (Automotive, Medical Imaging)

Francesca TRENTA

A dissertation submitted in fulfillment of the requirements for the degree of doctorate in Computer Science

Supervisor:
Prof. Sebastiano BATTIATO

Co-Supervisor:
Eng. Dr. Francesco RUNDO

Academic Year 2020 - 2021 (XXXIV cycle)

Abstract

In light of the tremendous success gained by Deep Learning algorithms, the role of these techniques is becoming increasingly important in the challenging automotive and healthcare fields. The proposed dissertation collects all the research works done by the Ph.D. candidate, focusing on the development of advanced Deep Learning approaches for the mentioned domains. The contribution of this dissertation is not only to deliver effective Deep Learning solutions for a wide range of problems but also to outline the challenges encountered in these fields. Broadly, the scientific community of both areas has always faced the lack of high-quality datasets, which has affected the performance of DL algorithms as they directly depend on data used to perform predictions. In the context of the automotive industry, Deep Learning has opened new opportunities to process large amounts of complex data coming from multiple automotive-compliant sensors. In fact, current literature is focusing on the designing of advanced driving support functions to analyze the car driver's physiological status for assessing his/her fatigue level. However, collecting physiological data posed several challenges due to the expense in terms of time and resources, which has hampered the research investigations. Hence, the datasets used in such studies are often quite limited. Furthermore, the driving scenario has increasingly pressed the development of real-time DL applications in order to provide a faster response. Despite the effort to support this demand, run Deep Learning pipelines in real-time driving scenarios still remains an open issue. Motivated by these issues, in Chapter 2, we attempted to define effective solutions to overcome the limitations previously mentioned, copying with the adverse conditions. A common problem in the automotive and medical imaging domain is related to the studies assessment, which must follow strict protocols devoted to preserving the privacy of users along with their safety. This process has hindered the collection of high-quality datasets, especially in the medical imaging field. Medical imaging has always suffered by the limited availability of properly labeled data and the scarce quality data (without noise or artifacts). In addition, clinicians are always rather skeptical towards Deep Learning approaches, as deep networks do not often lead to understandable results because they operate as "black-box," which do not make *interpretable* the computations performed among their intermediate layers. In Ch. 3, we reported a fairly comprehensive discussion related to the mentioned drawbacks. Although we attempted to enhance the current literature by designing promising Deep Learning approaches, we also focus on treating the issues to be overcome by researchers, enlarging the discussion related to the main challenges encountered nowadays in this field. Finally, in the Appendices section, we reported our findings in the aerobiology and quantitative finance domains, where Deep Learning approaches have been applied for solving many complex tasks.

Acknowledgements

First of all, I would like to express my gratitude to my Supervisor, Prof. Sebastiano Battiato, for valuable comments and discussions, support, and inspiration in the course of my Ph.D. journey. Without his support, this dissertation would not be accomplished.

Since this dissertation has covered several research domains focusing on the automotive industry and medical imaging branches, I would like to thank my Co-Supervisor, Eng. Dr. Francesco Rundo. His knowledge about these fields helped me a lot in fostering my thinking and attitudes.

I am also thankful to Dr. Alessandro Ortis for his supervision and timely directions that steered me towards the completion of my research works.

I would like to convey my sincere gratitude to Dr. Lorenzo Ascari and the "Ferrero" group, which have contributed to the collaborative projects of pollen grains microscope classification, especially providing a substantial amount of microscopic image data and manual annotations along with valuable comments.

I would like to thank Dr. Daniele Ravì, Senior Lecturer at the University of Hertfordshire, who has always been available to clarify my doubts and share me with plenty of inspiring ideas in the project of multi-modality classification.

This dissertation not only is the result of my Ph.D. course but also of the entire journey in Computer Science that began several years ago. A path filled with joyful times but also, and especially, with tough ones. For this reason, I would like to express my gratitude to the people who have always been to my side. In particular, I would like to thank my fiancé, Calogero, who has always believed in me when I couldn't even believe in myself. Without his support, I would not have achieved many goals. I share my deepest gratitude to my parents, who sacrificed everything to allow me to pursue my university education. I have never taken their efforts for granted. Last but not least, I would thank my sister, Laura, for her encouragement and support throughout this journey. Without her unconditional love, I would not have been able to reach here.

Contents

List of Figures

List of Tables

List of Publications

ADAS+

- *Patent*
  - Rundo, F. and Trenta, F. and Conoci, S. and Battiato, S. "Image processing method and corresponding system." U.S. Patent Application No. 16/729, 879, 2020.
- *International Journals*
  - Rundo, F. and Conoci, S. and Spampinato, C. and Leotta, R. and Trenta, F. and Battiato, S. **Deep Neuro-Vision Embedded Architecture for Safety Assessment in Perceptive Advanced Driver Assistance Systems: The Pedestrian Tracking System Use-case**. In: *Frontiers in Neuroinformatics*. 15, pp. 36, 2021.
- *International Conferences*
  - Rundo, F. and Conoci, S. and Trenta, F. and Battiato, S. **Car-Driver Drowsiness Monitoring by Multi-layers Deep Learning Framework and Motion Analysis**. In: *Sensors and Microsystems: Proceedings of the 20th AISEM 2019 National Conference (Vol. 629, p. 169).* Springer Nature, 2019.
  - Trenta, F. and Conoci, S. and Rundo, F. and Battiato, S. **Advanced motion-tracking system with multi-layers deep learning framework for innovative car-driver drowsiness monitoring**. In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, pp. 1–5.
  - Rundo, F. and Trenta, F. and Leotta, R. and Spampinato, C. and Piuri, V. and Conoci, S. and Labati, R. Donita and Scotti, F. and Battiato, S. **Advanced Temporal Dilated Convolutional Neural Network for a Robust Car Driver Identification**. In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*. LNCS (Vol. 12668, pp. 184-199), Springer International Publishing, pp. 184–199.
  - Rundo, F. and Conoci, S. and Battiato, S. and Trenta, F. and Spampinato, C. **Innovative Saliency based Deep Driving Scene Understanding System for Automatic Safety Assessment in Next-Generation Cars**. In: *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*. IEEE, pp. 1–6, 2020.
  - Rundo, F. and Spampinato, C. and Battiato, S. and Trenta, F. and Conoci, S. **Advanced 1D Temporal Deep Dilated Convolutional Embedded Perceptual System for Fast Car-Driver Drowsiness Monitoring**. In: *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*. IEEE, pp. 1–6, 2020.

  - Rundo, F. and Spampinato, C. and Conoci, S. and Trenta, F. and Battiato, S. **Deep Bio-Sensing Embedded System for a Robust Car-Driving Safety Assessment**. In: *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*. IEEE, pp. 1–6, 2020.
  - Battiato, S. and Conoci, S. and Leotta, R. and Ortis, A. and Rundo, F. and Trenta, F. **Benchmarking of Computer Vision Algorithms for Driver Monitoring on Automotive-grade Devices**. In: *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*. IEEE, pp. 1–6, 2020.

Medical Imaging

- *International Journals*
  - Rundo, F. and Banna, G. L. and Prezzavento, L. and Trenta, F. and Conoci, S. and Battiato, S. **3D Non-Local Neural Network: A Non-Invasive Biomarker for Immunotherapy Treatment Outcome Prediction. Case-Study: Metastatic Urothelial Carcinoma**. In: *Journal of Imaging*, 6.12, p. 133, 2020.
- *International Conferences*
  - Rundo, F. and Banna, G. L. and Trenta, F. and Spampinato, C. and Bidaut, L. and Ye, X. and Kollias, S. and Battiato, S. **Advanced Non-linear Generative Model with a Deep Classifier for Immunotherapy Outcome Prediction: A Bladder Cancer Case Study**. In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*. LNCS (Vol. 12661, pp. 227-242), Springer International Publishing, 2020.
  - Pino, C. and Palazzo, S. and Trenta, F. and Cordero, F. and Bagci, U. and Rundo, F. and Battiato, S. and Giordano, D. and Aldinucci, M. and Spampinato, C. **Interpretable Deep Model For Predicting Gene-Addicted Non-Small-Cell Lung Cancer In Ct Scans**. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*.
  - Rundo, F. and Banna, G. L. and Trenta, F. and Battiato, S. **Advanced Deep Network with Attention and Genetic-Driven Reinforcement Learning Layer for an Efficient Cancer Treatment Outcome Prediction** In: *2021 IEEE International Conference on Image Processing (ICIP)* (accepted)

Other Publications:

Pollen Classification

- *International Conferences*
  - Battiato, S. and Ortis, A. and Trenta, F. and Ascari, L. and Politi, M. and Siniscalco, C. **Detection and classification of pollen grain microscope images**. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 980-981, 2020.

- Battiato, S. and Ortis, A. and Trenta, F. and Ascari, L. and Politi, M. and Siniscalco, C. **Pollen13K: A Large Scale Microscope Pollen Grain Image Dataset**. In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 2456–2460.
- Battiato, S. and Guarnera, F. and Ortis, A. and Trenta, F. and Ascari, L. and Siniscalco, C. and De Gregorio, T. and Suárez, E. **Pollen Grain Classification Challenge 2020: Challenge Report**. In: *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*. LNCS (Vol. 12668, pp. 469–479), Springer International Publishing, 2020.
- Trenta, F. and Ortis, A. and Battiato, S. **Fine-Grained Image Classification for Pollen Grain Microscope Images**. In: *The 19th International Conference on Computer Analysis of Images and Patterns*. Virtual Event, Sept. 27 - Oct. 1.

Quantitative Finance

- *International Journals*
  - Rundo, F. and Trenta, F. and Di Stallo, A. L. and Battiato, S. **Advanced markov-based machine learning framework for making adaptive trading system**. In: *Computation*, 7.1, p. 4, 2019
  - Rundo, F. and Trenta, F. and di Stallo, A. L. and Battiato, S. **Machine learning for quantitative finance applications: A survey**. In: *Applied Sciences*, 9.24, p. 5574, 2019
  - Rundo, F. and Trenta, F. and Di Stallo, A. L. and Battiato, S. **Grid trading system robot (gtsbot): A novel mathematical algorithm for trading fx market**. In: *Applied Sciences*, 9.9, p. 1796, 2019

Chapter 1

Introduction

With the rapid development of technologies in the field of Artificial Intelligence (AI), data analysis has attracted much research attention over the last few years. However, the scientific community has immediately have dealt with typical issues related to extracting discriminative features from continuous data streams [1]. The increasing emergence in developing innovative pipelines to elaborate data signals has led to the spread of AI methods for extracting relevant features from such data. In this regard, Machine Learning (ML) algorithms, ranging from traditional ML methods to more advanced solutions such as Deep Learning (DL), provided a remarkable contribution. The main advantage of DL methods is the capacity to automatically learn meaningful features from a high volume of data, rather than traditional ML solutions, which involve the design of hand-crafted features. However, the effectiveness of these techniques directly depends on the amount and quality of available data. As stated previously, the data acquisition may generate some issues that negatively impact the quality of measurements. Despite achieving impressive results, current DL technologies have seen the general experts' skepticism in specific branches of research, such as automotive and medicine, which have slowed down the adoption of such approaches [2]. Hence, the collected data are currently validated and interpreted by human operators, representing an expensive effort in terms of cost and human effort required. In this regard, the scientific community has attempted to draw meaningful conclusions by providing new technically concrete and feasible solutions, exploiting advanced methods.

In the last few years, there has been a growing interest in analyzing human data in the automotive and healthcare domains. Researchers have provided evidence that an altered human physiological condition may induce dangerous situations (especially in a driving scenario) due to human processing failure and a slow reaction time [3]. Studies have pointed out that physiological alterations are often accompanied by physiological changes involving the human organs and tissues, such as the brain, heart, blood flow, and facial expressions. In the medical domain, much research attention has focused on processing clinical data for a wide range of applications, ranging from radiology image analysis to examination quality control. Performing clinical examinations of individuals could detect some biological anomalies, warning more severe health problems, and giving enough notice to assess an adequate clinical treatment or diagnosis [4], [5]. In this thesis, we investigate promising approaches related to the mentioned fields by analyzing physiological and medical imaging data, also discussing the problems encountered in our research activity. In these contexts, the research community is facing the following challenges:

- The difficulties correlated to acquiring and analyzing physiological data;
- The limited resources of automotive-grade devices' hardware;
- The non-availability of high-quality datasets;

- The lack of interpretability of DL models

This dissertation tackles these issues by building efficient and accurate DL-based systems for processing physiological and medical imaging data. As discussed, those challenges mainly arise from the difficulty of elaborate data sets, analyzing physiological signals, and acquiring large-scale labeled training data. Motivated by these issues, we provide a more in-depth study regarding the pros and cons of applying DL methods in two different research areas where data properties may hurdle the application of such advanced techniques: the automotive and medical imaging domains.

1.1 Automotive Field

In today's market, the automotive industry is continually evolving to improve the reliability and safety of the latest technology inside the cars. In order to satisfy the high demand for efficient and safe automotive systems, the technology is becoming increasingly sophisticated and the products currently on the market include intelligent solutions in the ADAS field. The ADAS systems, an acronym of *Advanced Driver Assistance Systems*, are becoming a very useful resource for the design of latest generation cars in order to increase the overall safety of driving, as well as to address classical automotive issues related to a drop in driver attention. Recent studies have pointed out that the driving assistance systems must be customized to the driver and his driving dynamics, profiling the driver or the recognition of his identity (both before and during driving) in order to enable the most appropriate ADAS assistance services. For this reason, there is a continuing demand to develop effective warning systems, not only to detect the presence of obstacles from the exterior environment, but also to deliver early detection of a car driver's inadequate physiological condition. The main objective is to minimize the risk of road accidents. Indeed, recent studies established that a considerable percentage of crashes are related to the bad driver status [6]. As a result, the most recent solutions involve the use of advanced DL algorithms for the analysis of physiological signals such as PhotoPletysmoGraphy (PPG) and Electrocardiography (ECG) [7]–[11] along with Computer Vision approaches devoted to processing head pose estimation, eye blink detection. In this context, ADAS technologies based on DL methods have a crucial role for the vehicle industry because they allow algorithms to automatically analyze driver's physiological signals to detect inattentiveness, drowsiness, road rage, and other potentially dangerous statuses [12], [13]. DL has achieved high levels of precision and accuracy in interpreting individual parameters extracted from physio measurements. However, the high accuracy comes at the price of large computational costs. Consequently, dedicated hardware devices, from the application of specific processors, are needed to optimize complex workloads of the DL methods. For this thesis, we discuss key challenges and core issues surrounding the study of a subject's physiological state using the collected signals from automotive-grade sensors. The main challenge is to address issues of having information affected by artifacts and noise. Further, we underline the main challenges of extracting physiological information from electrical bio-signals, such as PhotoPlethysmoGraphy (PPG) signals. In particular, we address the problem of detecting car driver's drowsiness. In the medical field, the "Drowsiness status" denotes a physiological state that leads to a reduction of the level of awareness and a tendency to fall asleep.

1.2 Medical Field

Analyzing the massive amount of information from individuals has become a topic of interest for researchers from the automotive field and the healthcare one. In the healthcare industry, clinical data comes from hospital records, medical examinations, such as Computed tomography (CT), and digital devices that can help monitor users' habits to gather information about their health condition. With the spread of advanced technologies, there is a strong demand to automate the analysis of clinical data to define suitable treatment guidance. However, this data requires proper management and analysis in order to derive meaningful information. In the last few years, the scientific community has focused its research efforts on delivering innovative strategies by analyzing visual features from medical images. Extensive collections of medical images applied to train deep neural networks have gradually become a hot topic in the research community and the medical industry. However, the image analysis process is often time-consuming and requires expert evaluations. Even simple tasks, such as image classification or segmentation, require a lot of expert clinicians' effort. Hence, there is a strong need to develop automated solutions for medical image analysis, especially in the medical oncology field. These considerations underlie the effort to study and develop new efficient algorithms to analyze medical images efficiently. More recently, the growing interest in DL technologies has led to the development of several approaches to accomplish these tasks. Recently, DL approaches have achieved outstanding performances in 2D and 3D medical image analysis tasks [14], [15]. In addition, it has significantly surpassed traditional visual feature extraction methods in a variety of tasks, including medical image analysis, by learning task-specific features directly from data. However, high performance in terms of accuracy is not enough in a safety-critical context such as the medical one. Indeed, two more aspects of crucial importance are *robustness* (to ensure the system will work even in case of perturbed data) and *interpretability* (to explain model decisions to physicians). This dissertation tackles both problems by proposing advanced DL pipelines devoted to performing segmentation and classification tasks. In addition, this research underlines the challenges of acquiring patient data in a clinical environment. Collecting health data poses several issues related to the missing labeled data and irrelevant feature selection. We address the mentioned problems using effective DL methods to handle the limited datasets, aiming to provide novel approaches. We also enforce *robustness* and *interpretability* in our proposed solutions, which is essential to make artificial intelligence accepted in clinical practices [16], [17]. Indeed, DL models show some vulnerabilities to perturbed data that may compromise their effectiveness and trustworthiness. In this regard, we report the existing defense techniques based on adversarial examples to improve DL model robustness, which leads to better interpretability.

1.3 Overall Structure

The dissertation structure is shown in Fig. 1.4. This work is organised as follows:

Chapter 2. Automotive: ADAS+. We investigate innovative approaches to detect bio-signals without wearing such medical devices or sensors, without imposing some driving positions on the car driver as they are impractical and not compatible with automotive standards. In particular, we propose efficient DL pipelines

designed to perform the non-invasive acquisition of physiological signals. Moreover, we indicate innovative solutions to assess both the driver's drowsiness by using physiological signal features and the correlated blood pressure level, providing a robust evaluation of the driving safety. Furthermore, we formally introduce the key challenges encountered in the analysis of physiological data when intending to assess the psycho-physical state of the driver. Although researchers have made many efforts to design safety assessment applications based on detecting physiological signals, embedding them into a vehicle environment represents a challenging task. These considerations furnish the motivation for showing the problems of existing solutions and their practical limitation in the automotive field. These issues are specifically concerning the acquisition of bio-signals susceptible to motion artifacts generated by body movements. We also report the limitation of using the invasive methodologies for the car driver that is often not feasible in the automotive field.

Chapter 3. Medical Imaging. We provide an overview on human data analysis in the medical field from several perspectives. We address the most critical issues related to the analysis of medical data, discussing how unlabeled datasets may hinder the application of DL models in this field. We detail the DL methods used in our research works for processing clinical data effectively. Firstly, we propose an effective method for improving the DL model robustness, thus enhancing the model *interpretability*, which represents a relevant property in the clinical context. Secondly, we focus on methodologies for extracting discriminative features in order to perform image modality classification task. Finally, we present our works related to the segmentation of infectious diseases, where several challenges have been addressed.

Chapter 4. Findings, Limitations and Perspectives. We summarize the thesis while providing a detailed discussion related to our findings. We also draw conclusions pointing out the proposed solutions and challenges encountered in our research investigation. Finally, we give ideas for possible future extensions of the presented research work.

Appendix A Deep Learning for Pollen Grain Microscope Images. We provide an overview of pollen object segmentation and classification in the field of aerobiology. We address the most important issues related to the standard palynological procedures relying on the manual classification of pollen grains by observing morphological traits on microscopy images. In particular, we discuss the hard work required by the techniques currently used to identify and count the relevant entities in microscopy that have hindered the application of aerobiology. Finally, we provide ML methodologies adopted to perform segmentation and classification of pollen grain detected in microscope images from aerobiological samples. We also defined a large-scale dataset composed of more than $13,000$ pollen grains. To the best of our knowledge, it represents the first available dataset with a large number of pollen objects.

Appendix B. Deep Learning for Quantitative Finance. We briefly summarize the challenges encountered in the financial field. In addition, we only provide insights related to our research works in this field as they are not directly related to the topics of this thesis.

1.4 Contributions

This thesis improves the state of the art in several directions:

- **Contribution 1**: We designed new approaches based on Computer Vision (CV) and Deep Learning (DL) techniques to evaluate the driver's status by using physiological signals in the automotive field. In particular, we proposed advanced solutions that support the use of DL approaches on automotive- compliant devices, alongside reducing the computational cost to develop a valuable tool for the automotive industry.

- **Contribution 2**: We explained the development of innovative pipelines to analyze medical images coming from different resources. In this respect, we investigated the problem of performing medical data classification with different modalities, which present heterogeneous characteristics determined by the tools used to acquire them. In addition, we discussed effective solutions for the segmentation of COVID-19 disease in clinical exams, which has become a pivotal point in medical research over the last year. Specifically, we adopted DL approaches based on pre-trained models to deal with the lack of accurately labeled data. We also discussed the concept of data interpretability. In this context, we indicated effective solutions based on "defense techniques," such as adversarial training, to enhance the model *interpretability*.

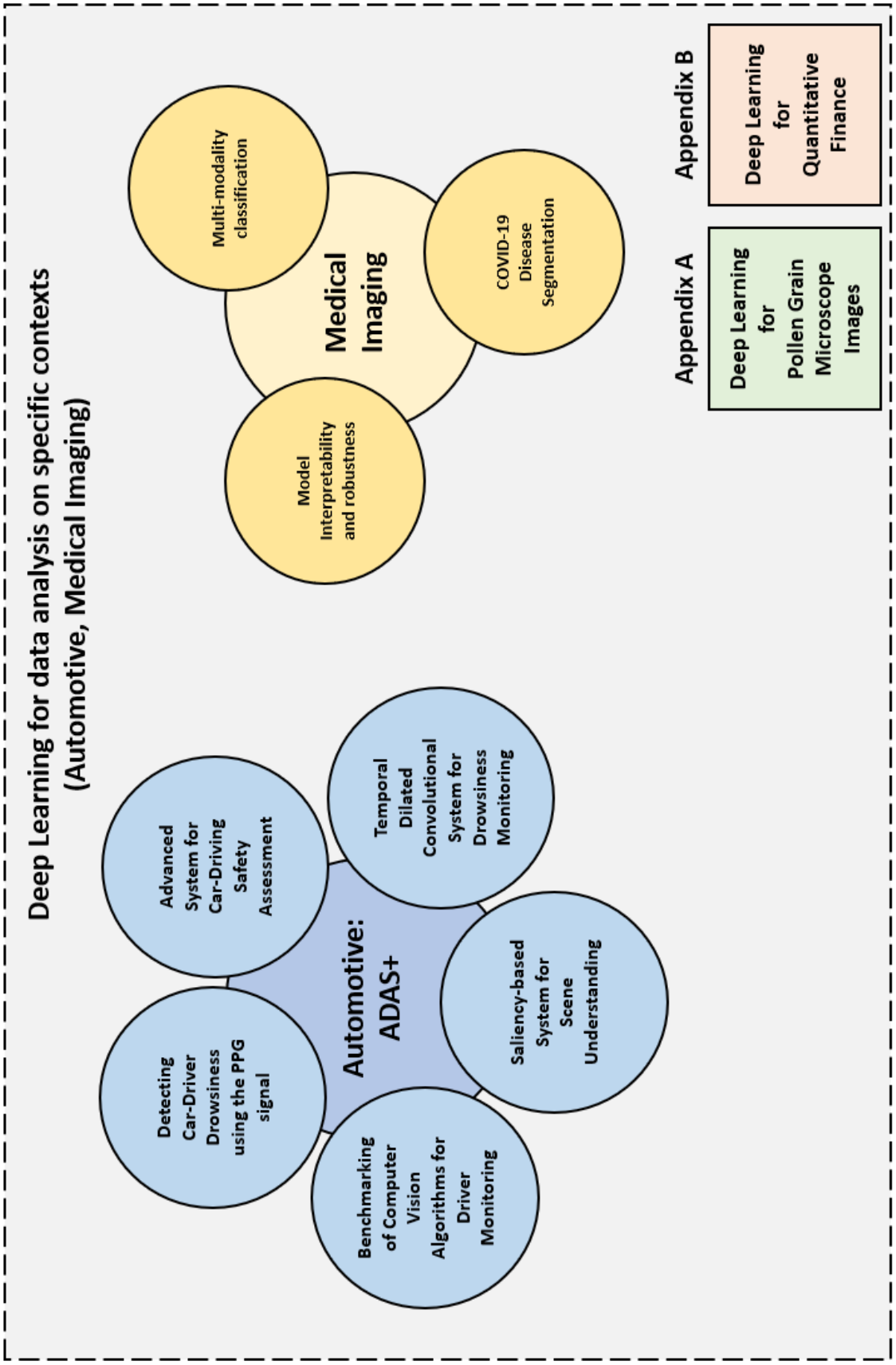


Figure 1.1: Overall structure of the current dissertation. Blue blocks represent algorithms for ADAS+ applications, whereas yellow blocks represent DL methods devoted to medical imaging analysis. Green block refers to algorithms for the analysis of microscopic images. Pink block indicates the implemented methods for Quantitative Finance field.

Chapter 2

Automotive: ADAS+

2.1 Overview

In the last few years, advances in technology have led to significant innovations in different fields. These changes are strongly linked to the automation of particular tasks that previously required human expertise. In discussing automation, we denote those activities performed using the term "currently demonstrated technologies" without requiring human intervention [18]. The concept of "currently demonstrated technologies" includes significant factors, such as technical feasibility, development costs, benefits from labor-cost substitutions, etc. Technical feasibility is a precondition for automation, whereas benefits derived from replacing human labor represent a crucial point for substantial reductions in cost (in terms of expense and time). Moreover, automation reduces the chance of human error. In fact, human performance is often error-prone, which consequently leads to misconfigured systems. These premises contributed to the development of automated objects, i.e., "smart objects". Nowadays, automation is transforming sectors such as healthcare and the automotive industry. In particular, the latter is deeply linked to the concept of "smart objects." With the recent digital transformation and the growing adoption of "smart objects," underlying the concept of the Internet of Things (IoT), the idea of designing high-tech cities has become a growing topic among researchers. The term "smart cities" refers to a sustainable, efficient, and innovative city, which provides a high quality of life to citizens thanks to developing advanced solutions and systems that are increasingly connected and integrated. The underlying idea of the IoT encompasses the interconnection of intelligent objects, such as smartphones and tablets, within the context of smart cities. Although developing "smart cities" poses some challenges, researchers have investigated effective solutions to realize their smart city vision. In the smart city vision, a wide array of applications, services, and technologies can connect a vehicle to other vehicles (V2V), to the infrastructure, and its surroundings. A connected vehicle encompasses a range of embedded features that enables car connectivity with other vehicles or infrastructure surrounding the vehicle, providing a more comfortable and safer driving experience. Specifically, a connected vehicle includes interactive "Advanced Driver Assistance Systems" (ADAS). The ADAS system comprises the following features:

- An electronic control system that adjusts the cruise speed of a vehicle, usually referred to as Adaptive Cruise Control (ACC). It also sets a specific speed that the car must maintain during the journey, slowing down or accelerating as needed.
- An efficient ADAS function consists of emitting a warning signal when a collision with another vehicle may occur. This system allows the car driver to be

warned in real-time. The warning system also warns of the presence of pedestrians.

- An embedded system for automatic braking in case of emergency which uses appropriate sensors and cameras to detect other vehicles or obstacles. If the system detects an obstacle, it acts on the brakes to decelerate the vehicle to avoid a collision.
- Another function of ADAS systems detects road signs using sensors and cameras. Specifically, the control system is able to recognize vertical, horizontal, complete, and luminous signs, as well as the speed set by the signs in order to warn the driver not to exceed this limit using an onboard display or by emitting a sound/visual signal if the limit is likely to be exceeded.
- A further ADAS function monitors the direction of travel and the lines of the carriageway in order to warn the driver if there is a lane invasion using a sound signal. In addition, this function makes it possible to keep the car within the limits of the lane through correcting the trajectory by acting on the steering.
- Another function of ADAS systems uses embedded sensors and cameras to facilitate assisted parking by alerting the driver to the distance between cars or other objects.
- There is also an LED headlight adjustment system that automatically changes the intensity level depending on the light conditions.

In recent years, smart vehicles have been equipped with warning systems for alerting drivers when unacceptable psycho-physical conditions occur. More specifically, the intelligent system provides an acoustic alert if the driver has a low level of awareness, thus avoiding dangerous situations. The automotive industry's main goal is not only to provide a smooth and adaptive travel experience but also to have a low environmental impact with zero emissions and to ensure zero fatal accidents. The goal of reducing road fatalities is a fundamental part of autonomous driving systems. Numerous studies have underlined the dramatic statistics regarding fatal road accidents. Several reports[1] outlined that approximately 1.35 million people die in road crashes each year. In particular, these statistics increase exponentially in reference to people aged between 15 and 44 years. In this regard, numerous efforts have been made by the researchers to provide valuable solutions, implementing advanced systems that avoid dangerous situations due to inappropriate driver behavior. Recent studies [19], [20] have also revealed that one of the main factors that leads to a major risk during driving is the unacceptable fatigue level. However, several research studies have outlined that distraction also plays a key role. Cutting-edge technology has allowed us to monitor both the internal and external vehicle environment while driving with the support of cameras that monitor the driver, thereby preventing possible dangerous situations. To this end, today's research has developed different approaches to assess the psycho-physical state of a driver, one of which is typically termed "drowsiness." This term identifies a state of a low level of awareness and inclination to sleep. Significant advantages in technologies have prompted researchers to develop effective methods to detect the critical level of driver drowsiness and avoid serious road traffic accidents. Assessing drowsiness represents a complex task that requires the evaluation of a subject's psycho-physical condition

[1] https://www.asirt.org/safe-travel/road-safety-facts/

in a short period of time when the situation is likely to degenerate, or the subject is showing hints of drowsiness. Detecting early symptoms of an inadequate physiological state may avoid possible accidents.

In this regard, researchers investigated the most promising approaches for detecting the driver's fatigue condition. Previous works focused on assessing the driver's physiological state by extracting visual information through the usage of CV techniques for tracking eye movement and head position. However, these approaches depend on the car lighting conditions or the high length of time to process all the data. Indeed, the computation time is the critical point, which is not feasible for real-time applications. Nowadays, the most adopted solution is based on the use of physiological signals such as ElectroCardioGraphy (ECG) or PhotoPlethysmGraphic (PPG).

A large number of studies have deepened the correlation between the level of attention and Heart Rate Variability (HRV) [21]. HRV is an index of autonomic control of the heart. HRV is obtained from frequency analysis of the ECG or PPG signals. Specifically, HRV reflects the heart beat-to-beat intervals that result mainly from the dynamic interaction between the Autonomous Nervous System and the heart. Indeed, the Autonomous Nervous System (ANS) activity is associated with heart rate activity. Within this perspective, recent scientific studies [21] analyzed the correlation between the drowsiness state and the ANS activity through the HRV.

Generally, existing approaches based on HRV detection propose invasive ECG sampling. The ECG signal raises a specific well-known issue related to the the lack of robustness of its sampling system in automotive applications. In fact, the ECG sampling requires at least three electrodes to be in contact with the human skin according to the minimum configuration known as Einthoven's Triangle [22]. Both car driver's hands must remain on the steering wheel where the two electrodes of the ECG signal sampling system are usually placed. In addition, the third electrode is placed on the driver's seat.

Considering this issue associated with the use of the ECG signal, scientists have investigated more effective solutions. In this regard, the most adopted solution consists of using the PPG signal. The PPG is a non-invasive estimation of blood volume changes in tissue by measuring characteristics of light either absorbed or reflected from tissue [23]. More specifically, PPG is a physiological signal generally acquired by using an optical sensing system that monitors the blood flow dynamic related to the cardiac phases [10], [23], [24]. PPG is a convenient and simple physiological signal that provides information about the cardiac activity of a subject [23] and, therefore, the drowsiness status of a subject as well as pathologies which may indirectly have an impact on the subject's guidance. In order to collect the PPG signal, the region of interest of the subject's skin is required to be illuminated by using a Light Emitting Diode (LED). By coupling the LED with a photo-sensing device, we sample the part of back-scattered light (photons), which is not absorbed by the blood flow of the subject on which the optical sensing system is placed [25].

However, physiological signals involve the use of automotive-compliant devices and processing data to provide a rapid assessment of the driver's physiological condition. For this reason, several works exploit ML and DL-based methods to analyze physiological signals consisting of many data points. The integration of devices for ADAS functions represents a delicate issue in that it is required by legislation that the ADAS system complies with different standardization protocols that regulate the technical specifications of the technologies adopted for its implementation. In this regard, several protocols place a significant emphasis on the safety provided by ADAS systems, often leading to difficulties when installing specific tools in the car.

Despite the difficulties that have emerged, automotive research has invested many resources in trying to find solutions to ensure a high level of safety and comfort while driving.

In this dissertation, we propose innovative methods to assess the driver's physiological state. This work was developed as part of the *ADAS+ project*, funded by the "PON Ricerca e Innovazione 2014-2020", which involves industrial partners such as STMicroelectronics Srl, and MTA Spa, along with the University of Catania and Chieti-Pescara. The *ADAS+ project* aims to develop an innovative demonstrator of safe driving assistance (ADAS+) capable of monitoring, in a timely and continuous manner, both the physiological level of the driver as well as his/her state of intoxication and the quality of the air in the vehicle. Silicon technologies, advanced image processing algorithms and nano-structured materials are integrated into a common platform that meets the safe driving standards required for the next-generation of "smart" cars. In particular, the project consists of the development of three advanced prototype technological modules, based on innovative technological platforms, such as:

- Fisio module: It consists of miniaturized silicon optical probes based on Silicon Photomultiplayer (SiPM) technologies integrated in the steering wheel. It monitors the level of attention (drowsiness) of the driver through the continuous control of the rhythm of the heartbeat and its variability.
- Vision Module: It comprises of (a) a Visible Light Micro Camera and (b) an Infra-red (IR) Light Camera to detect signs of fatigue, (c) a Silicon Radar/Lidar Device to detect obstacles outside the passenger compartment.
- Chemical Sensors Module: It consists of (a) Multichip with electrical transduction for monitoring the driver's level of sobriety/drunk state integrated into the steering wheel; (b) Environmental sensors microchip for monitoring the quality of the air in the passenger compartment using nanostructured materials such as Silicon NanoWires and Metal Oxides (MOx).

The technologies introduced in this dissertation will anticipate the not-yet-mature autonomous driving solutions, which ensure the possibility of increasingly safe vehicles in the future. Specifically, we have been involved in the development of the Vision module. Therefore, in the next sections, we will outline the application of CV and DL algorithms for automotive applications.

2.2 State of Art

A large number of literature has examined the most promising techniques for detecting driver's drowsiness. The outcomes have highlighted that these approaches could be subdivided into three separate macro-categories, reported as follows:

2.2.1 Behavioral parameter-based techniques

A large amount of studies have pointed out the relationship between eye blinking and fatigue [53]. In particular, eye movements have been identified as one of the best indicators of sleepiness. Early approaches were devoted to detecting user's drowsiness levels by exploiting CV techniques to analyze facial movements that characterize signs of fatigue. In particular, the approaches that analyze eye movements focus on how often eye blinking occurs, i.e., how often the driver closes his eyelids, which

Year	Paper	Input Data	Parameters
1997	Ogawa et al. [26]	Blink duration statistics	Behavioral
2003	Bergasa et al. [27]	Head position, lips, eye locations, face rotation	Behavioral
2006	Park et al. [28]	Eyelid closure	Behavioral
2005	Wang et al. [29]	Yawning	Behavioral
2007	Vural et al. [30]	Facial expressions	Behavioral
2007	Batista et al. [31]	Facial features and PERCLOS	Behavioral
2010	Friedrichs et al. [32]	Eye-tracking, blinking rate, PERCLOS, EYEMEAS	Behavioral
2012	Lenskiy et al. [33]	Skin-segmentation, eye blinking, eye closure	Behavioral
2007	Hong et al. [34]	driver's face, gaze, emotions	Behavioral
2017	Zhenhai et al. [35]	steering wheel angular information	Vehicular
2017	Li et al. [36]	Steering Wheel Angles (SWA)	Vehicular
2012	Mcdonald et al. [37]	Steering wheel, eye movement	Vehicular
2011	Vicente et al. [38]	ElectroCardioGraphy (ECG) signal	Physiological
2012	Szypulska et al. [39]	LF/HF ratio from ECG frequency signal	Physiological
2019	Lee et al. [40]	the pattern of Heart Rate Variability (HRV)	Physiological
2018	Ryu et al. [41]	Photoplethysmographic (PPG) signal	Physiological
2014	Kurian et al. [42]	Pulse Rate Variability from PPG signal	Physiological
2013	Li and Chung [43]	FFT and Wavelet-based features	Physiological
2009	Park et al. [44]	Pulse Wave	Physiological
2016	Sari et al. [45]	PPG signal, LF, HF, LF/HF	Physiological
2011	Lee et al. [46]	PPG signal, driver's face	Physiological
2012	Lee et al. [47]	eye movements, PPG signal	Physiological
2017	Cheon et al. [48]	PPG signal	Phsyiological
2018	Choi et al. [49]	Visual features, PPG signal	Phsyiological
2021	Tonni et al. [50]	Yawning, blink duration,	Phsyiological
2011	Monte-Moreno et al. [51]	PPG signal	Phsyiological
2019	Slapnivcar et al. [52]	PPG signal	Phsyiological

Table 2.1: Summary of the most relevant publications designed to assessing driver drowsiness.

is an evident hint of drowsiness. The aim of these methods is to identify, through CV techniques, how many times this movement occurs and especially to discriminate whether this movement is linked to drowsiness or not. In general, this type of approach analyzes the information previously acquired by video cameras and then processes the data. In Table 2.1, we show a list of research activities related to behavioral parameter-based approaches that have been published over the years. One of the earliest works in this field is [26] and dates back to 1997. In this work, the authors investigated the problem of assessing driver's drowsiness by analyzing user's facial images. Specifically, an efficient algorithm was developed to compute statistics on blink duration to determine driver drowsiness. This work confirms the relationship between alertness and blinking behavior. In [27], the authors propose a robust framework for driver drowsiness detection by using a single color camera placed on the car dashboard. The system is composed of three main stages. The first stage is devoted to detecting some of the main facial structures (mouth and eyes) alongside estimating head position. Moreover, they detect the user's skin using a stochastic model, called the *Unsupervised and Adaptive Gaussian skin-color model* (UASGM). The model locates the user's skin on a color image by computing the cluster related to a color that is in common to each user. Finally, a semantic segmentation algorithm is applied to searching lips and sclera colors (sclera is the white membrane covering the eye). The second stage performs face tracking by estimating the lips and eye locations frame by frame. The authors assumed that the distance between the camera and the user's face would stay constant. Moreover, the authors evaluate the vertical and horizontal face rotation by calculating the vertical and horizontal projection of the distance between the face center and its initial position in n iterations . The last stage determines driver vigilance by computing eye blinking or head rotation. The

system alerts the driver if his/her eyes are closed for more than 2 seconds or if the blinking frequency is more than 0.5 Hz. An alert is given when the driver prolongs a vertical and a horizontal rotation of 20° and 30°, respectively, for over 3 seconds.

In [28], the authors developed a robust system to determine the frequency of eye blinking for detecting driver drowsiness. More specifically, they estimated PERCLOS, i.e., a measure that computes the percentage of eyelid closure over a given period of time. A warning alert is given when PERCLOS exceeds a given threshold value. Moreover, Park et al. [28] tackled the problem of detecting driver drowsiness, providing an approach that considers the difficult lighting scenarios. The authors use a camera with an infrared band-pass filter to capture the driver's face during the daytime. Considering the challenging scenario during the night or cloudy days when the light source is scarce, the authors equipped the car dashboard by adding two IR LEDs (Infrared Light-Emitting Diodes). In both cases, infrared images that depict eyes with dark pupils are generated. Therefore, pupils appear clear during various illumination conditions. In addition, an algorithm for illumination compensation is applied to maintain a constant distribution of the intensity of the eyelid and pupil areas, compensating for the light intensity in the surrounding areas of the eye. Finally, the authors apply a cascaded Support Vector Machine (SVM) [54] to perform eye verification. Specifically, an SVM model classifies eyes into open eye groups and closed eye groups. In the first case, if open eyes are detected by the open-eye classifier, therefore the eyelid movement is estimated. On the contrary, if the open-eye classifier does not detect open eyes, they are determined by the closed-eye classifier, which measures the eyelid movements. Finally, PERCLOS estimation is computed.

In 2005, Wang et al. [29] developed a system for assessing driver drowsiness by evaluating the level of fatigue, focusing on driver yawning. Specifically, the authors designed an efficient detector of the face region by using the method proposed by Viola and Jones [55]. The face detection is performed by using as input recorded video that captures the car driver's face. A Kalman filter is applied to improve face tracking over video frames, thus speeding up face detection in the driving video. Considering that yawning movement represents a typical sign of tiredness or fatigue, the authors designed a method to detect the mouth region. First, they deducted the mouth area in order to estimate its degree of openness. In particular, they formulated an equation defined by the distance between the two lip corners and the distance between the two lines running through the upper and lower lip boundaries. Second, they also combined two approaches in order to detect mouth area. Sequentially, the largest connected blob is selected as the mouth region. Moreover, a Gaussian model is used to perform lip detection in RGB space. Following a similar approach to previous works, the authors found the lip corners using the integral projection in the vertical direction, observing that the local variations are performed vertically. Moreover, they determined the line linking the two lip corners representing the orientation of the mouth. In the last stage, they computed the integral projection of the vertical difference of mouth region and the direction of mouth rotation to identify the two lines indicating upper and lower lip boundaries. After extracting this information, the mouth openness is determined. Yawning detection consists of evaluating the mouth openness over the video frames. Specifically, yawning movement is performed when mouth openness exceeds a given threshold for N sequential frames. The results conclude that the method detects yawning movement for test video under various illumination conditions.

Vural et al. [30] introduced a system for detecting facial expression in order to assess car driver's fatigue during real driving scenarios. The participants are required

to use a system that simulates a driving scenario. The authors use an efficient system, called CERT, for capturing 31 different facial movements, including eye blinking and yawning. After detecting faces and eyes through a boosting technique, a series of Gabor filters are applied to normalize images. Finally, they employed an SVM classifier to predict drowsiness from facial movements. Results confirmed the effectiveness of the method.

The author of [31] designed a framework to identify facial features along with applying an elliptical face modeling to assess driver's drowsiness. Specifically, the author proposed an anthropometric face model to locate facial features, such as mouth, eyebrows, and eyes. After detecting skin region by computing skin histogram model, an ellipse is modeled and used as starting point for facial features detection. The main advantage of this model consists in identifying facial feature points considering different face gaze orientations. Finally, pupils are detected in order to measure PERCLOS value. As stated, PERCLOS is a reliable measure to evaluate human drowsiness. In addition, yaw and pitch head rotation is used to measure drowsiness.

Friedrichs and Yang [32] exploited eye-tracking for measuring car driver's fatigue levels. In their work, the authors used Karolinska Sleepiness Scale (KSS) [56]. KSS is a common reference for evaluating drowsiness. In experiments, the authors used the output signals from the camera tracking eye movement. After performing a few pre-processing steps, including data synchronization and valid eye blinks detection, an 18 feature extraction from eye signals is performed for drowsiness detection. In particular, they focused on eye closure, blinking rate, PERCLOS, and EYEMEAS. The latter value indicates the mean square percentage of the eyelid closure rating. The authors used These features to compute the correlation between them and KSS measure. Finally, a standard classification using common Machine Learning techniques is defined.

A novel method based on color and texture segmentation was proposed by Lenskiy et al. [33]. The proposed method consists of a neural network to perform skin color segmentation that locates mouth, eyes, nose, lips, and eyebrows. Firstly, they performed iris detection by using Circular Hough transform. The main contribution of this work is the skin-segmentation procedure. Specifically, the authors implemented a robust and adaptive skin segmentation method to detect facial regions (e.g., eyes, lips, nose, eyebrows, etc.) by extracting Speeded Up Robust Features (SURF) facial features [57]. After performing segmentation, eye blinking frequency and eye closure are estimated by monitoring the positions of irises through time.

In [34], the authors implemented methods based on image processing to track the driver's face, gaze, and emotions to assess his inattentiveness and drowsiness. Despite being promising, the image processing performance is affected by the condition of the in-vehicle environment scenario (e.g., light condition, occlusions, etc.).

2.2.2 Vehicular parameters-based techniques

Another pool of solutions committed to determining driver's fatigue level includes works based on vehicle information analysis, such as lane information, steering wheel angle, vehicle speed, etc. In Table 2.1, we listed research works for evaluating driver drowsiness through extracting vehicular-based features.

Zhenhai et al. [35] proposed a novel system to detect driver's drowsiness by collecting the steering wheel angular information. During experiments, the authors also examined the facial expressions to subdivide the driver's state into two wakefulness and drowsiness.

In [36], the authors developed an online drowsiness detection system to assess driver's drowsiness by collecting data from Steering Wheel Angles (SWA). First, the authors extract the features of real-time steering wheel angles. Next, they process the time-series of SWA in order to linearize the features. In addition, the system computes the warping distance between the series of linear features of sample data. Finally, the system evaluates the drowsiness level according to a decision classifier that uses warping distance.

Mcdonald et al. [37] proposed a method to predict fatigue-related lane departures through the use of the steering wheel and a random forest classifier. In the first stage, the system extracts drowsy-related lane departures using a modified Observer Rating of Drowsiness (mORD) scale. Moreover, the system computer PERCLOS measurement by analyzing eye movement from video frames. Finally, the extracted features are fed into a random forest classifier. Results demonstrated that SWA is a more reliable metrics than PERCLOS value.

2.2.3 Physiological Parameter-based Techniques

Detecting symptoms related to a lack of awareness during driving has been widely investigated by researchers. In this regard, the scientific community discovered that the level of attention is strongly correlated to cardiac activity. In Table 2.1, we report a list of research papers that investigated the application of physiological signals for designing reliable ADAS functions.

The main functions of the heart are regulated by the Autonomous Nervous System (ANS). Specifically, the sympathetic and the parasympathetic nervous system, the two branches of ANS, regulate many cardiac mechanisms, which are reflected in the attentional state of a subject [58]. In the automotive context, physiological signals represent a relevant data source to assess a subject's physiological condition [59] [42]. The study of physiological signals has received much attention from the scientific community of the automotive industry. Specifically, the growing proliferation of non-invasive medical devices to collect physiological parameters has led to advanced new tools to be integrated into the vehicle environment. Although researchers have made a large number of efforts to design safety assessment applications based on the detection of physiological signals, they faced several issues concerning the acquisition of physiological data. In particular, previous works have adopted techniques to acquire ECG signals to perform the analysis of the so-called Heart Rate Variability (HRV) and determine the psychophysical condition of a driver.

The authors of [38] introduced an innovative pipeline for monitoring a car driver's drowsiness analyzing the ElectroCardioGraphy (ECG) signal alterations analysis, which may introduce noise and artifacts while measuring HRV. Indeed, they proposed a pipeline to perform ECG signal stabilization along with classification based on classical linear discriminant analysis.

In [39], the authors proposed a reliable approach for detecting fatigue and sleep onset. Specifically, the authors showed a method to discriminate activity, drowsiness, and sleep, taking into account the low (LF) and high frequency band (HF) ratio detected over the heartbeats (or R-R) tachogram computed from the ECG frequency analysis.

In [40], the authors used three types of recurrence plots (RPs) derived from the R-R intervals of the heartbeats to feed a Convolutional Neural Network (CNN) for the classification of drowsy/awake status. Specifically, the authors investigated the pattern of Heart Rate Variability (HRV) to monitor the car driver's drowsiness [40].

In this work, the HRV is collected by using ECG and PPG sensors, which inevitably introduce some artifacts. However, the acquisition of the ECG signal poses some challenges that may compromise its usefulness. According to the configuration of Einthoven's Triangle [22], the driver must keep the contact of three parts of the body on the corresponding electrodes in order to acquire the ECG signal correctly. Hence, this involves issues in ensuring the robustness of the ECG signal sampling system from which the HRV signal descends [22]. On this basis, several studies have recently proposed the use of PhotoPlethysmGraphic (PPG) signal to analyze a subject's physiological status rather than the ECG [8]. Contrary to the ECG signal, the PPG signal requires only one contact point (subject skin) to be sampled [8].

In this regard, Ryu et al. [41] used red organic light-emitting diodes (OLEDs) and organic photodiodes (OPDs) for collecting the Photoplethysmographic (PPG) signal for detecting the drowsiness level. The benchmark evaluation confirmed the effectiveness of the proposed flexible PPG sensor, reporting a higher performance in terms of accuracy than the standard PPG probe.

In [42], the authors implemented effective algorithms to detect driver's drowsiness using the Pulse Rate Variability (PRV), a non-invasive signal obtained from PPG acquisition. After denoising the raw PPG signal, the authors detected the peaks of the PPG signal accurately, providing a reliable method for assessing the drowsiness status.

In 2013, Li and Chung [43] proposed a hardware platform consisting of a PPG sensor, a microprocessor unit (MCU), a wireless transmitter, a smartphone, and a server PC to provide a real-time response using the wavelet transform of HRV signals over short periods. Specifically, the aim is to classify drowsy and alert driving events using a Support Vector Machine (SVM) classifier. In order to detect the two driving events, the PPG signal is divided into 1-min intervals according to PERCLOS measurements. Next, the authors performed feature extraction using Fast Fourier Transform (FFT) and a Wavelet Decomposition. Finally, they compared the classification accuracy achieved by extracted features. The results confirmed that wavelet-based features led to a better classification performance than FFT-based features.

Park et al. [44] designed a non-invasive and affordable drowsiness detection system based on acquiring the driver's pulse wave from the PPG signal. The proposed systems comprise two main parts: the *Sensing part* and *Analyzing part*. The first consists of a PPG sensor placed on the steering wheel. In addition, they installed a PC laptop in the passenger seat using a rack to collect the driver's pulse wave. The latter is a system devoted to processing the acquired signal from PPG sensors in real-time. Specifically, the system detects feature points from the driver's pulse wave to evaluate the driver's drowsiness. The classification is performed by setting a threshold value. The overall results confirmed the effectiveness of the proposed approach showing an 83% drowsiness detection rate.

In [45], the authors introduced a two-stage that uses a Wavelet Packet Transform (WPT) and a functional-link-based fuzzy neural network (FLFNN) to provide early detection of car-driver drowsiness. First, the authors find the maximal peaks of the PPG signal after computing the first derivatives of raw data. Peaks provide information regarding Heart Rate Variability (HRV). Next, the authors extracted the first feature parameters, i.e., LF, HF, and the ratio of LF/HF, using a wavelet transform packet. Finally, the classification is performed by using a functional-link-based fuzzy neural network classifier.

Lee et al. [46] developed a non-intrusive system for detecting driver's drowsiness. The designed system takes two sources of input: the PPG signal and frame

images depicting the car driver's face. Face region is detected by using the Principal Component Analysis (PCA) algorithm devoted to extracting facial features. The main part consists of a Genetic Algorithm (GA) that performs drowsiness classification considering eye region from sample frames and the PPG waveforms characteristics. The PPG signal is obtained by placing the thumb on the PPG sensor module.

A similar approach is proposed by [47]. In this work, the authors used eye movement and the PPG signal to fed a multi-classifier for evaluating the driver's fatigue level. The multi-classifiers strategy includes an artificial neural network (ANN), a dynamic bayesian network (DBN), a support vector machine (SVM), an independent component analysis (ICA), and a genetic algorithm (GA). In order to extract features from the PPG signal, the authors used a hardware system comprising a PPG sensors module placed on the steering wheel, a Smartphone device to capture the driver's face, a wireless transceiver CC2420, and an ECG sensor module.

In [48], the authors used the PPG signal for extracting the drivers' bio-data. Specifically, a Support Vector Machine (SVM) classifier is applied to determine drowsiness conditions.

In 2018, Choi et al. [49] designed a system based on Multimodal Deep Learning that recognizes both visual and physiological changes in the state of attention of the driver. More specifically, they used a deep learning framework consisting of Long Short-Term Memory (LSTM) to classify the driver's condition based on both visual and physiological data.

More recently, Tonni et al. [50] used multi-modal data for detecting driver drowsiness. Specifically, they applied a CNN to process data related to the car driver's behavior (e.g., yawining, blink duration, etc.). Then, they used a LSTM autoencoder to detect anomalities in heartbeats. Finally, they designed an advanced CNN to elaborate information regarding road signs and vehicle speed. Combining the output of the mentioned architecture leads to the evaluation of driver drowsiness.

Over the last years, researchers have investigated blood pressure to determine the level of attention, alongside analyzing physiological signals. For example, Monte-Moreno et al. [51] proposed a non-invasive approach to estimate the systolic and diastolic blood pressure by acquiring a PPG signal. In order to estimate the blood glucose level (BGL), systolic (SBP), and diastolic (DBP) blood pressure, in [51] a signal processing module is designed to extract features from the PPG waveform to be used as input data for a different machine learning algorithms. The results confirmed that Random Forest achieves better prediction estimation. Slapnivcar et al. [52] investigated the problem of detecting Blood Pressure (BP) using an ML-based architecture. The authors overcome limitations derived from cuff-based devices using PPG and a downstream Deep Neural Network (DNN).

2.2.4 Other approaches: Driving profiling and Blood pressure estimation

In [60] the authors proposed an interesting approach for profiling the car driver behavior, including identity recognition. Through the analysis of the key-pressed patterns, the authors were able to discriminate the identity of a specific driver with acceptable accuracy.

In [61] a system named *Driver Adaptive Vehicle Interaction System* was implemented and analyzed. The main modules of the aforementioned method are: the *Profile Management Module* , the *Driver Management Module* and the *Interaction Management Module*. In particular, the first module is able to handle the car driver's identity and correlated driver's driving characteristics. After collecting data for each

subject, the authors provided a custom user-adaptive interaction system suitable to profile the driving dynamic.

In [62] the authors performed a proper investigation about a model of human driving behavior and its main correlated issues. They described an interesting discussion about the principal human factors that might impact driving: age, gender, personality, anger, mental stress, distraction, and so on. The authors have implemented and analyzed different pipelines with reliable results. In [63] the authors designed a pipeline named "Sense Fleet" based on the output analysis of such specific smartphone's sensors to identify and profile the subject who is driving. The method achieved good results. However, it suffers from the limitations of similar methods that use devices external to the car, therefore problems of invasiveness and compatibility with the automotive systems of the car.

In [64] the author proposed an interesting car driver identity recognition based on the usage of a combined approach which includes machine learning and dynamic time warping methodology. By analyzing such physiological signals of the car driver (collected from specific bio-sensor embedded on the car systems), the author recognized the driver's identity with high accuracy. The pipeline herein proposed is an improvement of the one described in [8]. Specifically, the authors propose an approach for the car driver identity recognition based on the analysis of the "physiological imprinting" of the subject [8].

In [65], the authors introduced an interesting vision-based driver assistance system for scene awareness using video saliency analysis. The results reported that the proposed pipeline could detect how the driver's gaze was focused during driving. In [66], the authors collected the eye-tracking data of 40 subjects consisting of non-drivers and experienced drivers when viewing 100 traffic images. In particular, the authors proposed a solution to assess the drowsiness level and a monitoring system regulated by the information of the driving scenario.

In [67], the authors analyzed several Android smartphone embedded sensors and classification pipelines in order to characterize the car driver behavior. The authors proposed a driver profiling approach using such Machine Learning based algorithms. More in detail, in the survey [67] the authors have investigated several promising solutions based on the usage of Support Vector Machines (SVM), Random Forest (RF), Bayesian Network (BN), and Artificial Neural Networks (ANN). The final results confirmed that the accelerometer and gyroscope represent the most appropriate sensors to monitor driving behavior. In terms of machine learning architectures, they proved that the RF is the best performing pipeline, followed by ANN even if the performance of both is satisfactory and equivalent, varying from 0.980 to 0.999 mean AUC values [67].

2.3 The research works

Considering the enormous impact of technologies based on the application of DL techniques and their remarkable results for the recognition and detection of objects, entities, and events (e.g., actions) from images and video streams, in this dissertation, we focused on the use of DL algorithms to monitor the car driver's physiological status. In particular, we detail the usage of the DL pipeline in combination with CV approaches to gather information about driver behavior through face analysis. As previously mentioned, identifying fatigue driving by using frontal facial expressions has been widely investigated by researchers [26], [27], [53] . In this regard, we enhanced the research by introducing novel approaches based on the analysis

of visual information to assess a driver's physiological condition. We assumed that facial features from video streams contain unapparent temporal variations that hold information about physiological status of a subject. These variations include color changes and invisible skin movements due to the volume of blood flow in the vessels. Consequently, detecting visual points over the video frames may also provide information regarding the cardiac activity and, therefore, ANS activity. In 2012, Wu et al. [68] proposed a novel method to amplify skin movements which are usually invisible to the naked eye. First, the author applied a temporal filter to spatial frequency bands, previously created by decomposing the input video sequence. Then, they amplified the filtered spatial bands by a given factor α. The amplified signal is added to the original one in order to produce the output video. As a result, the amplification reveals the variation of redness as blood flows through the face. This study furnishes us the motivation for investigating novel approaches for the analysis of cardiac activity through extracting visual features. Specifically, the proposed solutions include the use of advanced pipelines to exploit data acquired by internal cameras in order to infer additional information about the driver status (e.g., weakness, level of attention), and entities detected in the vehicle environment (e.g., other vehicles, pedestrian, roads, etc.). The development of the CV module takes advantage of analyzing images captured under high light condition. Furthermore, we also analyzed high and low-level aspects of the technology stack involved, considering the hardware and software environment in which these technologies will be embedded. In the context of automotive field, we considered the limitations of the hardware resources of automotive-compliant devices compared to those usually required for the implementation of modern CV and DL algorithms, which usually rely on massive computational and storage resources. In fact, developing a robust, reliable, and efficient embedded system has become a crucial point in the automotive industry. The technological advances and expansion of vehicle technology have led to more sophisticated embedded systems for vehicle control. Embedded systems are commonly used in several applications, including infotainment and telematics, safety, and powertrain control. The recent electronic innovation has contributed to introducing the development of new features for different functionalities. Hence, the modern embedded systems include microprocessors capable of efficiently running CV algorithms for several applications. However, there are many concerns related to automotive-grade hardware resources that present some limitations in size, memory, power, cost, etc. In this dissertation, we provided some interesting solutions that balance all requirements and limitations to execute advanced algorithms for ADAS applications.

During the research activity, we investigated the problem of assessing driver's level of fatigue by designing approaches that can broadly subdivided into three macro-categories.

2.4 Detecting Car-Driver Drowsiness using the PPG signal

In this section, we describe some methods for fatigue detection using physiological signals. As previously mentioned, the acquisition of physiological data requires automotive-compliant sensors to collect information about the driver's psycho- physical state. However, previous studies have adopted rather intrusive tools that contrast with the demand for the development of more integrated ADAS functions i.e., functions that ensure not only safety, but also a certain degree of comfort during the driving experience. Our work proposes several solutions that attempt to overcome

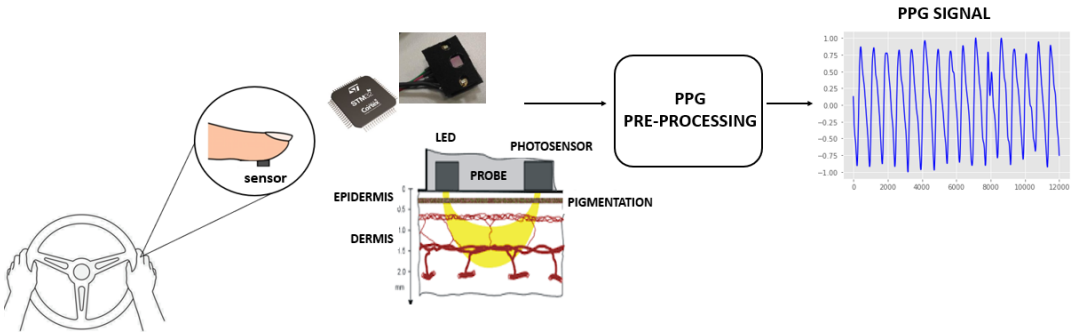


Figure 2.1: The PPG acquisition process.

the obstacles encountered in designing feasible solutions, developing methods that seek to minimize, as much as possible, the use of sensors and other tools within an in-vehicle environment [7], [69].

2.4.1 The PPG Sensing System

Before outlining the proposed approaches, we briefly summarize the characteristic of the PPG signal as well as the pipeline adopted for acquiring it. As previously mentioned, we implemented a system to acquire the car driver's PPG signal in order to monitor the attention level. The PPG signal is a non-invasive method for the analysis of the heart pulse rate. Indeed, PPG signal allows us to monitor both heart pulse and respiratory rate along with vascular and cardiac disorders [8]. Specifically, the PPG waveform consists of two components: a pulsatile 'AC' physiological signal that depends on the cardiac-synchronous changes in the blood volume, and a 'DC' component that shows minor changes due to the respiration and thermoregulation process [8]. When pumping blood to the periphery, the heart produces a specific pressure that distends the arteries and arterioles in the subcutaneous tissue. For this reason, we used a device composed of a light-emitter and a detector in contact with the subject skin to sample PPG signal. As heart pressure pulse can be seen as a peak in the PPG waveform, the changes in volume are detected by illuminating the skin and then measuring the amount of back-scattered light [8]. With regard to the PPG sampling, a device consisting of the Silicon Photomultiplier sensor was used [9]–[11] [24]. The PPG probes show an array detector device, called Silicon Photomultipliers (SiPMs) [11], with a total area of $4.0 \times 4.5\ mm^2$ and 4871 square microcells with 60 µm pitch. The devices have a geometrical fill factor of 67.4% and are packaged in a surface mount housing (SMD) with about $5.1 \times 5.1\ mm^2$ total area [70]. A Pixelteq dichroic bandpass filter with a pass-band centered at about 540 nm with a Full Width at Half Maximum (FWHM) of 70 nm and an optical transmission higher than $90 - 95\%$ in the pass-band range was glued on the SMD package by using a Loctite 352TM adhesive. The SiPM has a maximum detection efficiency of about 30% at 565 nm and a PDE of about 27.5% at 540 nm (central wavelength in the filter pass-band). As described, the PPG detector is composed of a light emitter in combination with a detector based on SiPM technology. The OSRAM LT M673 LEDs were used by SMD package and InGan technology [70]. The used LEDs devices are characterized by an area of $2.3 \times 1.5\ mm^2$ with a 120° angle view, a spectral bandwidth of 33 nm and a lower power emission (mW) in the standard range. To optimize the use of the PPG probe, a printed circuit board (PCB) consisting of a user-interface based on NI (National Instruments) instrumentation was designed. The PCB comprises a

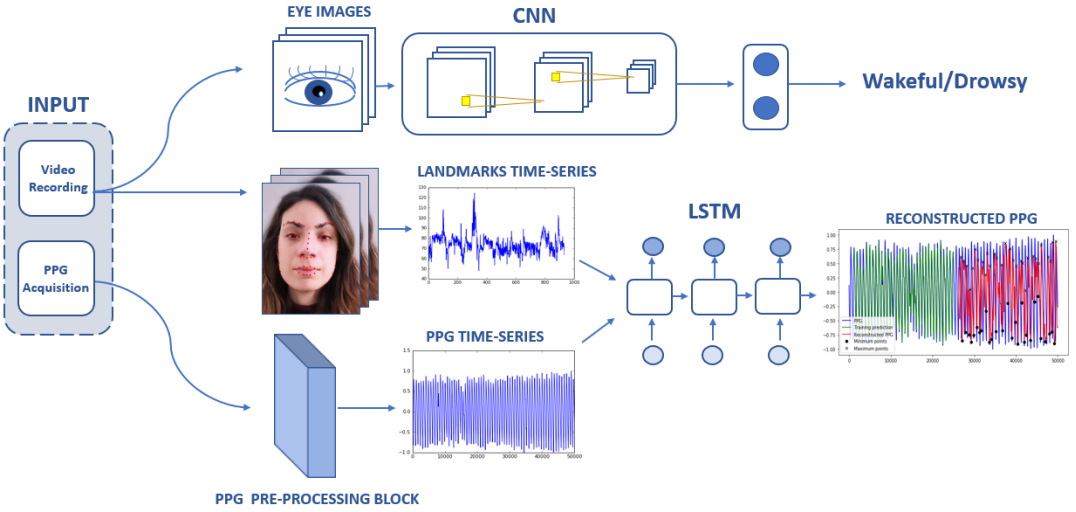


Figure 2.2: The overall pipeline.

4V portable battery, a power management circuits, a conditioning circuit for output SiPMs signals, along with several USB connectors for PPG probes and related SMA output connectors. In [9]–[11], [24], further details about the hardware used to acquire PPG signal were reported. The PPG Sensor Probe includes the SiPM sensor and the mentioned LEDs. The power consumption of the SiPM device is managed by Power management circuit [9]–[11]. We placed several PPG probes on the vehicle's steering wheel, in order to acquire the PPG signal. To collect the physiological signal, it is required that only one hand of the driver is placed over the PPG sensor. After collecting the PPG raw signal, we fed it into a national instrument (NI) framework consisting of several *Analog to Digital Converter* (ADC) with 24 bits of precision. Specifically, the NI device internally includes a Windows-based operating system with a LabView software framework devoted to performing some preliminary pre-processing of the PPG raw data [8]. More implementation details regarding the NI-LabView framework can be found on [8]. The procedure herein described generates a compliant standard normalized PPG waveform guaranteeing that the signal can be processed accurately. In Fig. 2.4.1, we illustrate the process underlying PPG waveform(s) formation.

2.4.2 The proposed pipeline

The proposed method comprises a mixed Long Short-Term Memory (LSTM) – Convolutional Neural Network (CNN) system [71]. In Fig. 2.2, we depicted the overall pipeline. The aim is to gather information about the car driver's drowsiness by calculating the HRV frequency domain of the driver's heart rate time series. Contrary to previous literature, our method consists of detecting and extracting facial landmarks using CV techniques to reconstruct the PPG signal instead of detecting it through the use of invasive sensors. The idea behind is based on the concept of Video Magnification [68] that can reveal tiny facial movements due to blood circulation and which can be challenging to observe with the naked eye. In order to take information about driver conditions, it is crucial to estimate the blood pressure that causes these movements by analyzing the HRV signal through detecting the PPG signal. With this aim, we extract facial landmarks from images of the car driver's face. The landmark points are used to gather information about fundamental facial components (e.g., nose, mouth, eyes, etc.) [72]. In order to extract these points, we

recorded a video capturing the driver's face. For each video frames, we detected the landmark's pixel intensities and their variations which then composed the landmarks' time series used as input for the LSTM pipeline. The advantage of the LSTM neural network is represented by its ability of detecting dependencies in sequential data (time-series). In particular, the LSTM understands long-term dependencies making it suitable for sequence learning tasks. Furthermore, we designed a CNN model to perform binary classification regarding the physiological states (wakefulness or drowsiness). The mentioned architecture is fundamental to validate LSTM results.

2.4.3 Experiments

In this section, we describe the experimental setup. A total of 71 subjects were involved in the data collection. We proceeded to collect 71 drivers aged between 20 and 70 years old, both male and female. Specifically, we collected healthy subjects and sick ones that have several pathologies, such as cardiac problems, hypertension, diabetes, etc. The data acquisition was conducted in accordance with the Helsinki Declaration of 1975. The study was approved by the local Ethical Committee of University of Catania. After signing the written consent form, subjects' frontal face was recorded using a color camera device having a max resolution of 2:3 Mpx and 40 fps as the framerate. The collected dataset includes the PPG signal measurements and video sequences of the subject's face. In particular, we performed the acquisition of the PPG signal considering two different states:

- State of Wakefulness: Under physiologists suggestions, we emulated a full wakeful scenario confirmed by simultaneously ECG signal sampling while recording, so as to show Beta waves indicating a high wakeful brain activity. In this context, subjects were asked to reply to specific questions in order to keep them engaged and active;
- State of Drowsiness: We emulated a full drowsy scenario confirmed by simultaneously ECG signal sampling while recording, so as to show Alpha waves - meaning a state of low arousal - under the supervision of a team of physiologists. Subjects were asked to perform eye blink movements while recording the video sequence to signal they were having a drowsy state.

Due to the preliminary nature of the study, we performed the experiments under high-light conditions in order to facilitate the video acquisition. Future studies will also address the detection of driver drowsiness under low-light conditions. Measurement sessions for each scenario were around five minutes long to ensure the preliminary system calibration and for real-time continuous learning. A similar setup was also used for the experiments detailed in the following paragraphs. As mentioned, a video sequence of each subject's face was recorded by using a low frame rate (25fps) Full-HD video camera. After this stage, we detected facial landmarks from video frames by applying Python's Dlib library that extracts facial features from images, i.e., nose, eyes, mouth, and face shape [72]. Finally, we computed the pixel intensity associated with each landmark point. The variation of point values over video frames defines the landmark time series used to feed the designed LSTM network.

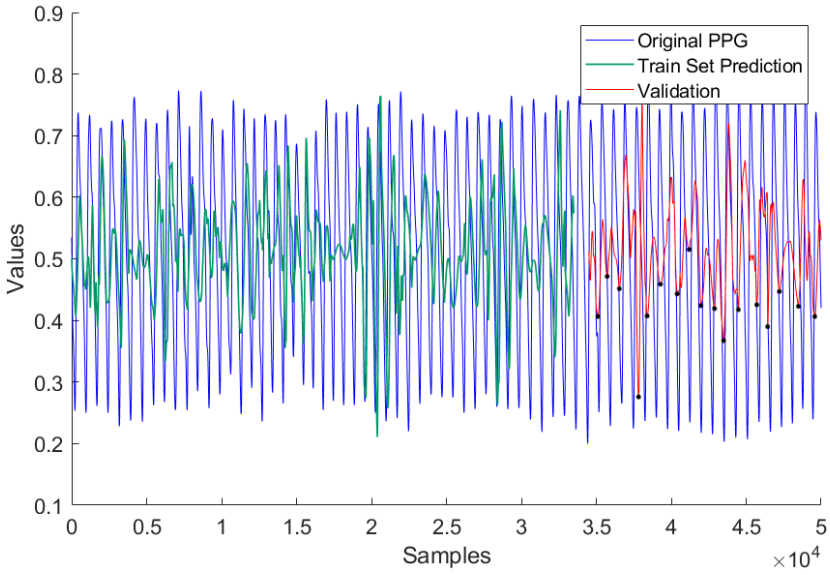


Figure 2.3: The minimum points of the reconstructed PPG signal (in red).

2.4.4 CNN Block

We implemented an ad-hoc CNN model to track the facial movements of the car driver. The aim is to improve the drowsiness detection robustness of the proposed pipeline. After acquiring the PPG signal for preliminary system calibration and real-time continuous learning, we used the classification results to validate LSTM prediction. The proposed CNN architecture is based on LeNet architecture [73]. The model training is performed using a batch of size 32, with an initial learning rate of 0.001. Moreover, we designed our model by using Adam optimizer. We defined 32 neurons in the hidden layers and two output neurons, as the target dataset has two classes. The experimental results reported that our model classifies the images accurately. In particular, we reached an accuracy of 80%, but some validation results reported an accuracy close to 90%.

2.4.5 LSTM Block

In this section, we detail the development of the proposed LSTM pipeline. First, we recorded a video sequence depicting the face of a subject. We used a low frame rate video camera in order to avoid the usage of intrusive tools. Next, we extracted video frames for detecting facial landmarks related to the main facial structures such as the nose, the mouth, and the eyes. In this regard, we adopted Python's Dlib library that is an implementation of the algorithm developed by Kazemi [72]. We collected the PPG signal using automotive-compliant sensors placed on the steering wheel, representing the target data. Facial landmarks make up the time series used to feed the LSTM neural network. All values were scaled in the range (0.2, 0.8) with the standard MinMaxScaler algorithm. The training is performed using 256 neurons, a batch of size 128, and an initial learning rate of 0.001. Moreover, to prevent overfitting, we used a dropout ratio of 0.2. Finally, we extracted the minimum points of the PPG sensor-based signal and the PPG reconstructed signal (from LSTM) in order to verify the robustness of the LSTM output. For this purpose, we computed the distance of these points to compare the distance between the PPG minimum points

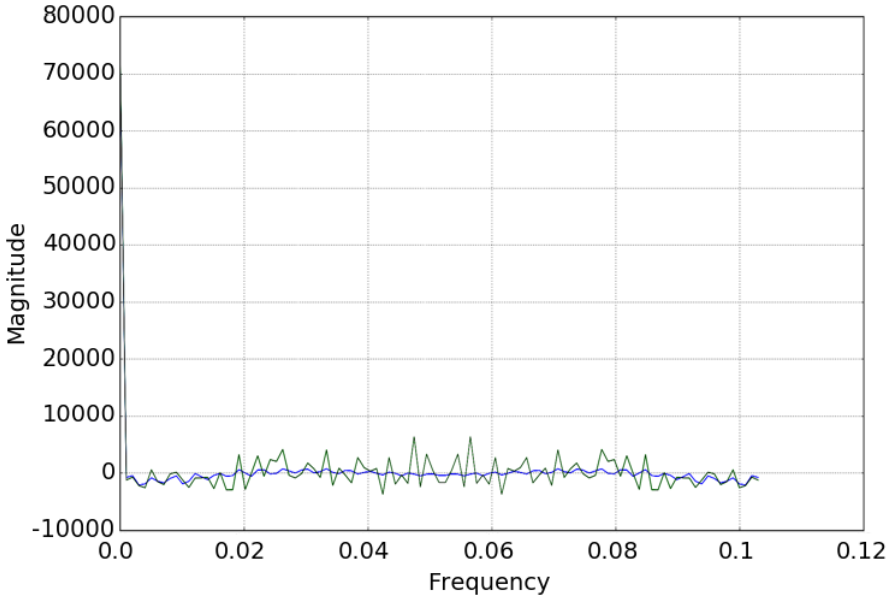


Figure 2.4: Correlation between FFT Spectrum of the original PPG minimum points (blue) and reconstructed PPG minimum points (green).

and the distance of the reconstructed PPG minimum points. In Fig. 2.3, the detected minimum points are shown.

2.4.6 Results and Discussion

In order to demonstrate the effectiveness of the proposed method, we computed the correlation between the Fast Fourier Transform (FFT) spectrum of the PPG sensor-based minimum points and the FFT spectrum of the reconstructed PPG minimum points. In Fig. 2.4, we provided the result of the FFT spectrum. Moreover, we tested our proposed approach using the reconstructed PPG time series for computing HRV dynamic to evaluate the drowsiness of a subject (car driver) included in our dataset. The results showed very promising performance since it is able to distinguish drowsy subjects from wakeful ones with high confidence reporting similar results with previous work in this field [74], [75]. We proposed a sensor-less method in order to obtain information about the drowsiness state of a car driver. The experimental results confirm a correlation between PPG collected through the usage of sensors and the PPG reconstructed by using facial landmarks. Our proposed method does not require invasive and expensive devices to record the PPG signal; we defined it from facial landmarks. We present a valid solution to overcome the main drawbacks of other works (e.g., the driver has to put his hands over the steering wheel to record the PPG signal).

The results presented in this section represent preliminary results considering the rather limited number of subjects composing the dataset, which did not allow us to perform a comprehensive study. However, we have shown that the proposed method is quite effective in detecting driver drowsiness and that it can be implemented, without expensive costs, within an ADAS system. Future work will focus on building a more diverse and numerous dataset to determine more robust results.

Given the results of that work, we extended our research to monitor a driver's drowsiness by applying different approaches that consider not only physiological information but also the vehicle's surroundings as well as the user's blood pressure to determine signs of fatigue. These studies are based on the previous pipeline, although they have been extended with additional pipelines to define more comprehensive work.

2.5 Advanced System for Car-Driving Safety Assessment

In this work [76], we expand the pipeline described in the section 2.4 to monitor driver drowsiness. The proposed approach attempts to define an effective solution for a driver's fatigue level using non-intrusive tools. To provide an additional degree of robustness to this investigation, we enlarged our prior work by developing a novel pipeline for the robust detection of drowsiness levels. In this respect, we developed a robust and non-invasive system to monitor the PPG signal and assess the subject's blood pressure, representing an advance in automotive monitoring systems.

2.5.1 Pipeline

The proposed Visio2PPG Reconstruction Pipeline

In this section, we describe the Visio2PPG architecture that comprises a PPG sensing device for preliminary system calibration coupled with a CV camera module to sample the subject's face for extracting facial descriptors. An ML algorithm that correlates the subject's face descriptors with the corresponding PPG waveforms completes the proposed pipeline. In the training calibration stage, the proposed pipeline correlates the time-dynamic (in all sampled video frames) of the selected face descriptors with the corresponding PPG features to evaluate the car driver's drowsiness and blood pressure level. During the calibration phase, the PPG signal is collected using the coupled LED-SiPM sensing device, described in Section 2.4.1. After removing motion and noise artifacts from the PPG time-series, we detected the minimum and maximum extreme points by computing the first and the second derivatives of the filtered PPG signal. Simultaneously, we recorded a video sequence of the subject's frontal face with a camera device in order to preliminary identify significant landmarks or descriptors on the driver's face, according to the strategy proposed in Section 2.4.

We analytically formalize the reconstruction of landmark dynamics by using Kazemi and Sullivan algorithm $\psi_{KS}(.)$. Formally, let be $I_t(x, y)$ the captured $M \times N$ gray-level (or luminance channel in case of colour camera device) driver face video frames at the instant t_k , the *i-th* landmark dynamic time-serie $l_i(t_k, x_l, y_l)$ is reconstructed using the following equation:

$$\ell_i(t_k, x_l, y_l) = \psi_{KS}(I_{t_k}(x, y));\ k = 1..N_f; i = 1..N_L \tag{2.1}$$

where $\ell_i(t_k, x_l, y_l)$ represents the pixel intensity dynamic of the *i-th* landmark identified at the frame position (x_l, y_l) while Nf represents the number of captured frames and NL represents the number of 68 identified landmarks (Eq. 2.1). Once detecting landmarks dynamics, we analyzed the descriptors time-series $\ell_i(t_k, x_l, y_l)$ in order to correlate their temporal intensity with the underlying cardiac activity. The main

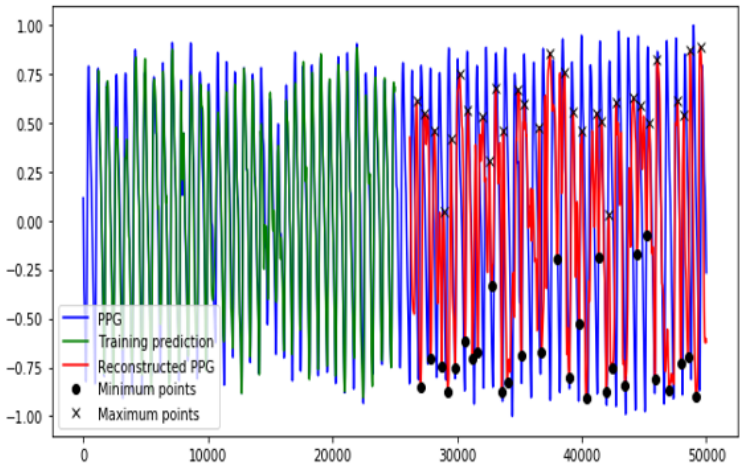


Figure 2.5: Minimum and maximum points of the reconstructed PPG signal.

advantage of the proposed method consists in not requiring the use of high-frame rate camera devices.

The Deep LSTM architecture

In this contribution, we significantly improved the approach previously described in [7] (Section 2.4) based on employing a Deep Long Short-Term Memory (D-LSTM) architecture to reconstruct some features of the PPG signal to characterize the driver's drowsiness level. Specifically, we proposed ad-hoc D-LSTM framework to correlate the driver's facial landmarks and the corresponding cardiac activity. The D-LSTM network is based on Vanilla architecture initially proposed by Hochreiter and Schmidhuber in [77]. The proposed D-LSTM architecture is composed of 1 input layer, 2 hidden layers, and 1 output layer. Specifically, we designed the input layer with 64 units and the two hidden layers with 64 and 128 cells, respectively. The output layer comprises 1 cell providing the predicted PPG sample. Each LSTM layer is followed by a batch normalization and a dropout layer appointed to boost the overall performance. We trained our model with an initial learning rate of 10^3. We also set the batch size to 512, and the maximum number of training epochs to 200. In training-calibration phase, the proposed Deep LSTM learns the correlation between the selected facial landmark time-series $\ell_i(t_k, x_l, y_l)$ with the corresponding sampled PPG signal, for each subject. The output of the so designed Deep LSTM pipeline represents predicted extreme points of PPG signal. Specifically, we perform the following computation:

$$\mu_\ell(t_k) = \frac{1}{N_L} \sum_{j=1}^{N_L} \ell_j(t_k, x_l, y_l) \tag{2.2}$$

Using Eq. 2.2, the so computed signal $\mu_\ell(t_k)$ will be fed as input of the D-LSTM. The deep architecture will be trained to correlate the input signal $\mu_\ell(t_k)$ with such features of the corresponding PPG signal of the analyzed subject i.e. with the extreme points *m1,m2,m3,m4*. Once the Deep LSTM framework has learned the correlation between the driver's facial landmarks and the extreme points of the corresponding PPG signal, the training-calibration phase will be dropped. The calibration phase requires 15/20 seconds of PPG signal (our sensing device performs acquisition at 1 kHz) and corresponding visual frames (at 40 fps). Finally, the underlying

trained model provides a feed-forward estimation of the PPG extreme points. In Fig. 2.5, the detected minimum and maximum points of the PPG signal are depicted. An HRV block completes the pipeline providing a corresponding attention measurement based on classical frequency analysis of the so determined extreme points of PPG [8].

The Blood pressure estimation pipeline

In this section, we describe the pipeline used to monitor the driver blood pressure since it is strongly correlated to the driving safety. We propose a novel solution to measure blood pressure level of the subject by simply acquiring the corresponding facial video frames and then processing it with the proposed Vision2PPG reconstruction pipeline. As described, we retrieved the extremal points of the subject's PPG signal. These points are fed as input of a properly configured Shallow Neural Network that classifies the normal-blood pressure subjects from those who report pressure values beyond the norm. After collecting the set of extremal points of the PPG waveforms, we characterized the subject's cardiac activity depending on the blood pressure. Formally, for each pair of PPG waveforms PPG^j, PPG^{j+1} , we define the following indicators:

$$\varphi = \left[m_1^j, m_2^j,\ m_3^j,\ m_4^j, dx_i^j, dy_i^j, mAI^J\right] \tag{2.3}$$

$$m_i^j = \left(x_{m_i^j}, y_{m_i^j}\right) \tag{2.4}$$

$$dx_i^j = x_{m_i^{j+1}} - x_{m_i^j}, \tag{2.5}$$

$$dy_i^j = y_{m_i^{j+1}} - y_{m_i^j} \tag{2.6}$$

$$mAI^j = ((y_{m_3^j} - y_{m_1^j}) - y_{m_4^j})/(y_{m_3^j} - y_{m_1^j}) \tag{2.7}$$

$\forall\ j = 1..(N^{PPG} - 1)$, $i = 1, 2, 3, 4$, where mAI^J is a modified version of the so-called Augmentation Index, usually computed for measuring the arterial stiffness while N^{PPG} represents the number of estimated extremal points (Eq. 2.7). The other indicators reported in the Eqs. 2.3 - 2.6 characterize cardiac cycles and, therefore, the corresponding blood pressure level. The elements of the vector φ represent the input of the aforementioned Shallow Neural Network (SNN) designed to learn the correlation between the computed input of elements and the corresponding value of the systolic and diastolic blood pressure. The designed SNN (with an hidden layer of 500 neurons) is trained with the Scaled Conjugate Gradient backpropagation (SCG) algorithm [78]. The output of the SNN framework is a binary value which is a discriminating flag denoting if the subject shows normal pressure values (0) or not (1). The set 120/80 indicates 120 mmHg for systolic pressure and 80 mmHg for diastolic pressure. Under the supervision of a team of physiologists, we setup 120/80

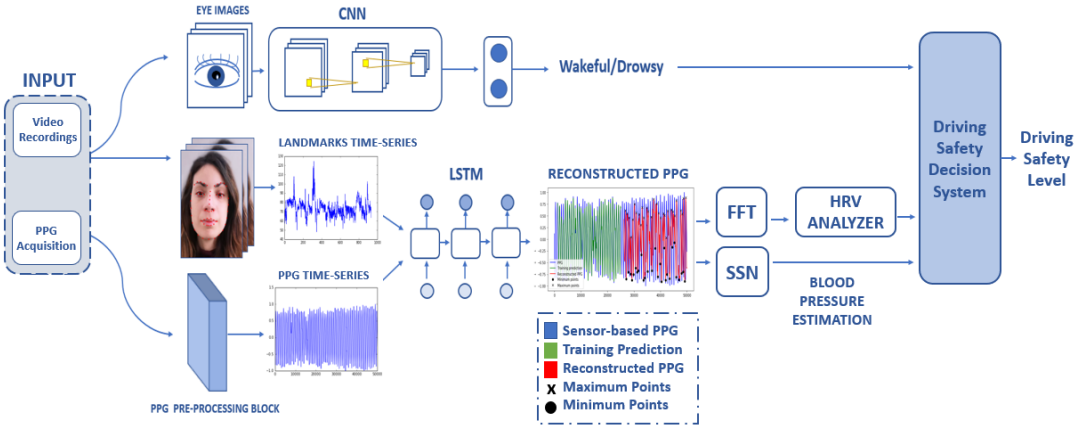


Figure 2.6: The overall scheme of the proposed pipeline.

mmHg as acceptable blood pressure whereas higher/lower values are considered anomalous.

The Deep CNN Pipeline

To assess the car driver's drowsiness, the authors propose an innovative pipeline based on ad-hoc Deep Convolutional Neural Network (D-CNN). The proposed D-CNN network is composed of three convolutional layers. Except for the last convolutional layer, each layer is followed by a ReLU activation layer (with batch-normalization) and 2×2 max-pooling layers. The first convolutional layer performs 32 operations with 3×3 kernel filters, where the second and the third ones show 64 and 128 kernel filters of 3×3, respectively. A Softmax layer completes the D-CNN pipeline performing binary classification of the visual input data i.e., drowsy/wakeful status of the analyzed driver. We performed fine-tuning for 100 epochs using Adam optimizer and cross-entropy as the loss function. To conduct our experiments, we also set the learning rate to 0.001 and the batch size to 32. The input visual frame of the D-CNN is a segmented car driver eye (77×77 resolution) obtained through Haarcascade classifier [55].

The driving safety monitoring

In this work, we proposed a pipeline that combines the driver's physiological and visual data, sampled by specific sensors placed in the car. Specifically, we implemented a pipeline that reconstructs the features of the driver's PPG signal from visual information when the steering embedded sensors will do not provide real PPG data. The extracted features allow us to reconstruct the driver's level of attention by analyzing HRV and evaluating the blood pressure dynamics providing a robust measure of the driving safety level. We increased the robustness of the proposed approach by processing the video frames of the driver's face through a D-CNN model, performing a further classification of the driver's attention level. In Fig. 2.6, the overall scheme of the proposed driving safety estimation architecture is shown. We have implemented a block called Driving Safety Decision System (DSDS) to analyze the outputs produced by the previous pipelines (i.e., HRV analysis from PPG or Visual landmarks based reconstructed signal, SNN output related to blood pressure estimation from PPG or Visual reconstructed PPG, D-CNN classification of the facial expression of the driver). In detail, DSDS generates an acoustic signal, whether at

Method	Drowsiness Estimation		
	Overall Accuracy	*Drowsy Driver*	*Wakeful Driver*
Proposed	95.07%	95,77%	94,36%
[24]	94,38%	95.83%	92.95%

Table 2.2: Car Driver Drowsiness Performance Comparison.

least one safety control block estimates a significant level of risk. The acoustic signal is managed by the STA1295 Accordo5 system, which hosts the DSDS software implementation. To sum up, the Vision2PPG pipeline reconstructs the extreme points of the PPG signal of a monitored car driver. Therefore, the distribution in the frequency domain of the reconstructed PPG extreme points ($m1, m2, m3, m4$) from the proposed pipeline is robust and allows to obtain a frequency spectrum (FFT) very close with the real one (i.e., determined by the driver's real PPG signal). Through this reconstructed PPG data, we obtained a robust HFT suitable to characterize the driver's drowsiness level [30]. Moreover, a PPG-based blood pressure analysis will be performed by the SNN sub-system, which is correlated to driver drowsiness. A further DCNN architecture works simultaneously with the Deep LSTM framework, providing a further assessment of the driver's attention level based on the analysis of facial features as illustrated in [7] (Section 2.4). The system provides a measure of the overall driver's attention level by analyzing the rankings of the three mentioned sub-systems..

2.5.2 Dataset

In this section, we describe the experimental setup. A total of 70 subjects both healthy (45 subjects having a blood pressure less or equal to 120/80 mmHg) and hypertensive (25 subjects having a blood pressure higher than 120/80 mmHg) took part in the data collection. The minimum age of the subjects was 20 years, while the maximum age was 75 years. After signing the consent form[2], subjects were recorded (frontal face) using a color camera device having a max resolution of 2.3 Mpx and 40 fps as framerate. The collected dataset includes systolic and diastolic pressure and PPG signal measurements, plus video sequences of the subject's face. Moreover, subjects performed different drowsiness states under the supervision of a team of physiologists. Simultaneously, we collected the EEG signal from subjects determining the real level of attention. Each measurement session is around 10 minutes long, 5 of which is in a state of full attention and 5 in a state of low attention. The PPG was acquired by using the described sensing device with a sampling frequency of 1 kHz. In order to collect the blood pressure measurements, we used a classic certified sphygmomanometer. In the collected dataset, the minimum pressure value is around 105/70, while the maximum pressure value is 160/95. The collected dataset was split into 70% for training, 15% for validation, and 15% for testing.

2.5.3 Experiments

In our experiments, a NVIDIA GeForce RTX 2080 GPU as used for training and testing. Moreover, MATLAB full toolboxes rel. 2019b environment for Deep Learning framework were employed. We compare our proposed method with another similar

[2]Ethical Committee CT1 (authorization n.113 / 2018 / PO)

Method	Blood Pressure (BP) Estimation		
	Overall Accuracy	*Class 1 Normal BP*	*Class 2 Abnormal BP*
Proposed	88.73%	88.88%	92.30%
[79]	88.73%	91.11%	84.61%

Table 2.3: Blood Pressure Performance Comparison.

pipelines [24]. Table 2.2 shows the accuracy benchmark comparisons both regarding to the early recognition of the low-attention driver (Drowsy Driver) and to the early detection of the high-attention driver (Wakeful Driver). A measure of the average performances is reported (overall accuracy). This comparison highlights the effectiveness of the proposed method, showing a 2% increase in accuracy with respect to [24]. We also validated the performance of SNN network used to estimate the blood pressure level by comparing our method with other approaches in literature [79]. Table 2.2 shows the BP values measured both to normal-pressure subjects (Class 1) and to those with an abnormal BP (Class 2). As confirmed by the metrics in Table 2.3, the margin of error for the implemented pipeline is also considerably low. The results for SSN network suggest that our model provides reliable classification results as confirmed by the overall accuracy of 88.73%.

2.6 Temporal Dilated Convolutional System for Drowsiness Monitoring

In this section, we detail our approach for evaluating driver drownsiness by combining an embedded time-domain hyper-filtering algorithm for the PPG signal and a 1D Temporal Convolutional architecture with a progressive dilation setup [80]. An extension of the detailed research activity was proposed on [81].

2.6.1 The Hyper Filtering Layers

The main contribution of this work is the filtering signal method used to stabilize the PPG raw data. In order to tackle the presence of motion and noise artifacts during the PPG acquisition, we performed a frequency filtering and a signal stabilization of the physiological raw data. Specifically, we employed a set of FIR filters to perform a low/high-pass and filter at the range of $1 - 10$ Hz, removing the 50 Hz power line frequency noise and other artifacts. In order to ensure robustness, we designed a bio-inspired algorithm to complete PPG signal stabilization [8]. Previous works have outlined the effectiveness of the mentioned bio-inspired pipeline [8], [10], [11], [25]. We enhanced the filters setup with another set of hyper-filtering layers based on hyperspectral imaging that collects visual information from the electromagnetic spectrum [82]. Hyperspectral imaging denotes a methodology to perform the frequency spectrum of each pixel of image [82]. In our study, we investigated the hyperspectral imaging considering 1D signals. We verified whether it is possible to retrieve information about the driver's level of attention by collecting the hyper-filtered information to more than one frequency range of the PPG signal. In particular, we analyzed a range of frequencies to characterize the value of the single PPG waveform. In this respect, we observed that the useful frequency range is included in the $1 - 10$ Hz range. In addition, we employed both a low-pass and high-pass

Type	Frequency pass [Hz]	Frequency stop [Hz]	Passband Attenuation [dB]	Stopband Attenuation [dB]
LP	4.8	10	0.001	100
HP	1	0.3	0.01	40

Table 2.4: Low-Pass and High-Pass Filter Design for the PPG Signal.

filtering. In our configuration, we performed two layers of hyper-filtering, which changes the frequencies in the low-pass part, maintaining constant the cut-off frequency of the high-pass filter (i.e., Hyper low-pass Filtering layer) and vice versa. In Table 2.4, we detailed the classical filter setup related to the raw PPG signal. We opted for the use of Butterworth filters in both layers of hyper-filtering in order to not introduce noise artifacts [24], [83]. We used a "try-and-error" approach to determine the number of sub-interval in the 1-10Hz range. After performing heuristic tests to detect optimal filter sub-bands, we reached a value equal to 11 sub-intervals. This number reflects the best trade-off between computational load and discriminative capacity. Therefore for each Hyper-filtering layer, we proceeded to subdivide the range of applicable frequencies (if low-pass or high pass) in 11 specific sub-bands. Once subdividing the frequency into 11 sub-bands, we designed a Reinforcement Learning (RL) algorithm. The implementation details of this approach are reported in the following items:

- We defined an action a_t as the sub-band frequency selected in the range reported in Table 2.4 and according to the type of filtering (low-pass or high-pass);
- an Agent is defined selecting the action a_t
- We defined a next state S_{t+1} as a set of pre-processed signals obtained collecting the value of each input PPG samples (in a windows of 5 sec sampling at 1 Khz as sampling frequency) of the filtered PPG raw signal at specific sub-band frequency of the action a_t;
- We define an environment Reward as $R(.|s_t, a_t)$ i.e., a measure of drowsiness of the car driver. We indicated as $R(.|s_t, a_t)$ the distance of the output of the deep learning system (regression layer plus SoftMax classification) with respect car-driver's level of attention.

We determined the optimal policy P_o that minimizes the cumulative discount reward by applying the following formula:

$$P_o = argmax_{P_o}\ E[\sum_{t\geq 0} \gamma^t R(.|s_t, a_t)|P_o\ ] \tag{2.8}$$

Where γ is a proper discounted coefficient in (0,1). In order to evaluate the the goodness of a state st and the goodness of a state-action couple (s_t, a_t), we denoted the Value function and the Q-value function respectively:

$$V^{P_0}(s_t) = E[\sum_{t\geq 0} \gamma^t R(.|s_t)|P_o\ ] \tag{2.9}$$

$$Q^{P_0}(s_t, a_t) = E[\sum_{t\geq 0} \gamma^t R(.|s_t, a_t)|P_o\ ] \tag{2.10}$$

By applying Q-learning algorithms [84], we determined a proper set of sub-band frequency for each hyper-filtering layer Eqs. 2.9, 2.10. In Table 2.5 and 2.6, we summarized our findings as results of the developed RL algorithm. Once identifying

F	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
HP	0.5	/	/	/	/	/	/	/	/	/	/
LP	0	1.4	2.9	2.5	3.8	3.9	4	4.5	5	5.3	6.9

Table 2.5: Hyper Low-Pass Filtering Setup (in Hz).

F	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
HP	0.5	1.2	2.6	2.7	3.3	3.5	4	4.4	5	5.7	6.4
LP	7	/	/	/	/	/	/	/	/	/	/

Table 2.6: Hyper High-Pass Filtering Setup (in Hz).

the frequency sets of the two hyper-filtering layers, we generated a set of filtered PPG signals. Specifically, we processed the raw PPG data by using the frequency sets reported in Table 2.5 and 2.6. Formally, let be $W^i_{PPG}(t_k)$ the single segmented waveform of each hyper-filtered PPG time-series (i.e. by using a specific frequency values both in low-pass and high-pass). We computed a signal-pattern for each sample $s(t_k)$ of the waveform, which depends on the variation of signal samples $s(t_k)$ in each of the hyper-filtered PPG time-series. In this way we collect a large dataset of signals obtained by dynamic analysis of each samples of each hyper-filtered PPG signal. The size of the dataset is equal to the number of filtering frequencies, i.e. 11 (see Table 2.5 and 2.6). In order to detect the car driver's drowsiness (from which we collect the PPG signal), we extracted a large amount of signal-patterns from hyper-filtered PPG time-series considering a timing window of 4 sec. Our designed Deep Learning block takes as input the extracted signal patterns. These signals will be used to characterize the driver's level of attention through the next Deep learning block.

2.6.2 The Deep Learning block

As introduced, we implemented ad-hoc Deep 1D Temporal Dilated Convolutional Neural Network (1D-CNN) [85]. Our proposed network takes as input the signal-patterns $s(t_k)$ generated from each PPG hyper-filtered signal. The main novelty of the proposed architecture is the introduction of dilated causal convolution layers. The term "causal" denotes that the activation at time t depend on the activation computed at time $t-1$. The proposed 1D-CNN includes such multiple residual blocks. Specifically, the proposed deep network consists of a sequence of 12 residual blocks stacked together. Each block consists of a dilated convolution layer, batch normalization, ReLU and a spatial dropout. A dilated layer includes a 3×3 convolution operation. The dilation factor is set to 2 increasing for each block. A downstream two-classes softmax layer completes the proposed pipeline. The output of the designed 1D-CNN predicts the driver's drowsiness level from source hyper-filtered PPG generated signal-patterns.

2.7 Results and Discussion

We evaluated our proposed pipeline on our dataset composed of PPG measurements from different subjects. For this purpose, 70 patients were enrolled. The age range of the recruited patients was $21-70$. To validate the effectiveness of our designed

Method	Car Driver Attention Estimation	
	Drowsy Driver	*Wakeful Driver*
Proposed	98.71%	99.03%
[24]	96.50%	98.40%

Table 2.7: Benchmark Performance of the Proposed Pipeline.

1D-CNN framework, we collected PPG signals emulating both drowsy and wakeful scenarios. Under the supervision of experts in physiology, we collected the PPG signal alongside acquiring the ECG signal in order to confirm the awareness of a subject. Previous studies have pointed out that EEG signals confirm the subject's level of attention by analyzing the presence of alpha and beta waves [9], also demonstrating the correlation between the physiological signal and the drowsiness level. Similar to the previous study [7], we collected the PPG signal of each recruited subject with a sampling frequency of 1 kHz. We collected the PPG signal in drowsy and wakeful conditions. We collected data for 5 minutes. To conduct the experiments, we used 70% of the total dataset for the training and 30% of the collected data for testing and validation. To validate the robustness of the proposed approach, we exploited ad-hoc PPG signals composed of parts of the signal acquired both in a high and low attention status. We process the so collected PPG-based signals to evaluate the robustness and efficiency of the proposed pipeline to discriminate the range of attention levels. In the Table 2.7, we reported the performance of the proposed pipeline compared with other similar pipelines based on deep learning approaches [24].

The classification results indicate that the proposed method is more effective than other works based on Deep Learning approaches. A major benefit of the proposed solution consists in not requiring the acquisition of the ECG or EEG to assess the driver's drowsiness level. As previously described, the sampling of these physiological signals is affected by motion and noise artifacts that compromise the HRV analysis. Moreover, another benefit relies on avoiding the PPG data analysis in the frequency domain as requested by other approaches based on HRV analysis. In our work, we also underlined the advantages of acquiring a PPG signal in a vehicle environment, positioning properly embedded sensors on the steering wheel. The experimental results also suggest that the proposed pipeline requires only about one minute of PPG acquisition to accurately classify the driver's attention level (with respect to $10 - 12$ min required by HRV based methods). Furthermore, the accuracy value could be improved if the implemented pipeline is combined with other promising solutions to recognize the level of attention, such as the imaging methods in the visible and infrared spectrum. The deep learning framework has proved to be adequate for learning data (signal patterns) generated by the pre-processing of the waveforms of the acquired PPG signal. Indeed, our results highlight the effectiveness of the designed pipeline in assessing the driver's drowsiness level. Specifically, the performance of the overall pipeline achieves reliable results on the collected dataset, which includes time-series of the collected PPG signal related to 70 recruited subjects. As reported, we acquired the PPG signal by emulating both drowsy and wakeful scenarios. The implemented Deep Learning algorithm is currently being ported to an embedded system based on the SoC STA1295 ACCORDO 5 produced by STMicroelectronics (software environment with embedded Linux), which is equipped with several ARM Cortex cores and a GPU for 3D graphics and a video and image dedicated pipeline [9], [79], [86]–[88]. Moreover, the pipeline to filter

and stabilize the collected PPG raw data is being ported to a hardware/software environment based on SPC5x CHORUS microcontroller technology released by STMIcroelectronics [89]. More discussion about the proposed pipeline is provided on [80].

2.8 Saliency-based System for Scene Understanding

The aim of the ADAS+ project is to create an innovative safety-driving system by analyzing the psychological state of a driver. In this respect, the research activity has been further extended to include not only approaches based on drowsiness detection but also methods to analyze visual saliency. Visual saliency is the human attention mechanism that encodes such visio-sensing information to extract features from the observation scene. In the last few years, visual saliency estimation has received significant research interests in the automotive field. While driving the vehicle, the car driver focuses on specific objects rather than others by deterministic brain-driven saliency mechanisms inherent perceptual activity. Moreover,

In this study [90], we propose an intelligent system that combines a driver's drowsiness detector with a saliency-based scene understanding pipeline. Specifically, we implemented ad-hoc 3D pre-trained Semantic Segmentation Deep Network to process the frames captured by automotive-grade camera device placed outside the car. Similar to the previous approach [80], we defined an ad-hoc 1D Temporal Deep Convolutional Network to classify the collected PPG time-series in order to assess the driver's attention level. Finally, we compare the detected car driver's attention level with corresponding saliency-based scene classification in order to assess the overall safety level.

2.8.1 Pipeline

The Drowsiness Monitoring System

In this work, we implemented a system to acquire the car driver's PPG signal to assess and monitor the related attention level. In order to collect the PPG signal, we placed several PPG probes on the vehicle's steering wheel. As previously mentioned, it is required that only one hand of the driver is placed over the PPG sensor. After acquiring the PPG raw signal, we performed a preliminary process to filter the PPG raw data. In this respect, we defined ad-hoc FIR (Finite Impulse Response) filtering both low-pass and high-pass band. Furthermore, we perform such preprocessing operation by computing the first and second derivatives of the collected PPG signal to evaluate the min and max value for each waveform. Finally, we rendered the PPG signal. Hyper-filtering approach is an innovative method devoted to processing a signal [24]. Similar to the previous work 2.6, we used hyper-filtering method to denoise the source PPG signal and process it with different frequency setup in order to extract discriminative patterns related to drowsiness level of a subject. Once the hyper-filtered PPG signal patterns [11] are collected, an ad-hoc 1D Temporal Dilated Convolutional Neural Network (1D-CNN) is used to classify the so generated patterns. The proposed Temporal Convolutional Network consists of a Dilated Causal Convolution layer that processes the time steps of a given sequence. Indeed, the term "causal" means that the activations computed at time t derive only from inputs from time $t-1$. The architecture of our proposed network is composed of multiple residual blocks, each containing two dilated causal convolution layers

Figure 2.7: Saliency analysis of the video representing the driving scene.

with the same dilation factor, followed by normalization, ReLU activation, and spatial dropout layers. More specifically, we implemented a 1D-CNN of 12 blocks, including a downstream softmax layer. Each block structure consists of a dilated convolution layer with a 3×3 kernel filters, a spatial dropout layer, followed by another dilated convolution layer, a ReLU layer, and a final spatial dropout. In this context, we set the initial dilation size to 2, increasing it at each block. Finally, a softmax layer completes the proposed pipeline. The output of the network is a scalar number in the range [0,1] that indicates the attention level of the car driver, discriminating drowsy (0) and wakeful drivers (1).

The Video Saliency Scene Understanding Block

In this section, we provided architectural details of the proposed Deep Network to perform saliency analysis.

We designed an ad-hoc 3Dto2D Semantic Segmentation Fully Convolutional Network (SS-FCN) to process the captured video frames, depicting the driving scenario. The SS-FCN architecture is mainly composed of two parts: an encoder and a decoder. The encoder (3D Enc Net) is devoted to extracting spatio-temporal features. It is composed of 5 blocks. Convolutional layers are gathered in 2 blocks of 2 layers for the first two blocks. The kernel size of each separable convolutional layer is $3 \times 3 \times 3$. Each block includes a sequence of two $3 \times 3 \times 3$ convolution operations, which is followed by a batch normalization, a ReLU layer and a max pooling operation with a pooling size of $1 \times 2 \times 2$. This sequence is repeated two times. Then, a sequence of $3 \times 3 \times 3$ convolutional layers is defined. Similar to the previous blocks, the succession of two convolution operations is repeated for the remaining three blocks, followed by a batch normalization, and an another convolutional layer with $3 \times 3 \times 3$ kernel, a batch normalization and a ReLU with a downstream $1 \times 2 \times 2$ max-pooling layer. The decoder backbone (2D Dec Net) is designed to decode the visual features. It consists of 5 blocks, and up-sampling layers. It also includes 2D

convolutional layers having 3×3 kernel. After each convolutional layer, we employed batch normalization layers and a ReLU layer. The residual connections are added through convolutional blocks. The decoder part is implemented to adjust the size of feature maps by up-sampling operations. The output of the proposed SS-FCN architecture is the feature map of the acquired video frame related to the segmented area of the most salient object in the driving scenario. In semantic segmentation, usually the fixation points comprise the most salient objects of the entire image. In Fig. 2.8.1 we depicted such instances of the SS-FCN output reporting the driving scenario saliency maps.

The Driver Attention Analyzer

This block estimates the overall driver's attention level, coming from the PPG signal analysis combined with the output (i.e., saliency map) of the proposed SS-FCN block. Specifically, this block verifies that the level of attention is coherent with the "meaning" of the corresponding salience map i.e. with driving scenario dynamic. A static saliency map reflects a low dynamic in the driving scenario, involving a low-level attention, whereas an increasing variation of the salience maps require a car driver's full attention. Formally, let be $S_f(x, y, t)$ the saliency map and O_S the output of the 1D-CNN, the designed block analyses the driver attention by using the following equation:

$$V(S(x, y, t)) = \begin{cases} \frac{\partial S_f(x,y,t)}{\partial t} \leq \vartheta, & O_S \leq \varphi \\ \frac{\partial S_f(x,y,t)}{\partial t} > \vartheta, & O_S > \varphi \end{cases} \tag{2.11}$$

Using Eq. 2.11, the Driver Attention Block checks the dynamic changes of the saliency map $S_f(x, y, t)$ through a given threshold ϑ. We set the ad-hoc threshold φ to verify the level of attention obtained by the Drowsiness Monitoring systems. Both thresholds are determined during the training phase by means of a heuristic calibration i.e. by choosing a setup that maximizes the performance of the overall pipeline during the learning.

2.8.2 Dataset

We evaluated the implemented pipeline on the DH1FK dataset [91]. We also collected video sequences of several driving scenarios using a camera with a 2.3 Mpx resolution and a maximum framerate of 60 fps. Furthermore, we recruited subjects showing different levels of attention while collecting the corresponding EEG signal [9] to confirm their physiological status. The acquisition was performed under the supervision of a group of physiologists. The used dataset was collected under the clinical study Ethical Committee CT1 authorization n.113 / 2018 / PO. To conduct our experiments, we involved a total of 43 subjects in the age group of 21-70 years. We acquired the PPG signal of the subject with a sampling frequency of 1 KHz, considering a 5-minute interval. Moreover, we used 70% of the all acquired PPG time-series and video driving scene frames for training, the remaining 30% for testing and validating. To sum up, the overall proposed pipeline assesses the driver's attention level by combining the PPG signal and the saliency map.

2.8.3 Experiments

The The SS-FC was evaluated on DHF1K dataset [91], using the Area Under the Curve (AUC), Similarity, Correlation Coefficient and Normalized Scanpath Saliency

performance indexes. Results suggest that the proposed architecture achieves acceptable performance (AUC: 0.885; Similarity: 0.355; Correlation Coefficient: 0.455; Normalized Scanpath Saliency: 2.564) compared with similar architectures [92]. In addition, results outlined that the architectures with a better performances, require a high computational cost as well as the use of invasive hardware. In this regard, our proposed pipeline shows low workload without involving specific hardware accelerations [88].

As previously mentioned, the Driver Attention Analyzer block compares the level of attention required by the driving scenario with the output of the Drowsiness Monitoring System. The system alerts the driver with an acoustic signal if the driver's vigilance level is lower with respect to the current driving scenario. In this context, we defined two range of values to classify the attention level. A value (output of the 1D-CNN based Drowsiness Monitoring System) in the range of 0 - 0.6 indicates a medium-low attention level. On the contrary, a high attention level comprises the values from 0.61 to 1. Finally, we set an ad-hoc normalized threshold ϑ (0.45) to define a static scene-based saliency map (see Eqs. 2.11). The driving scene is considered as 'dynamic' if the values of the normalized saliency map gradient are greater than 0.45 (requiring a high level of attention). For a normalized saliency map gradient lower than 0.45 (requiring a low level of attention) the driving scene is considered 'static'.

2.8.4 Conclusion

The encouraging results have confirmed that the proposed pipeline could assess the driver's drowsiness to preserve driving safety. The main benefit of the proposed method is that it does not require frequency domain analysis, compared to other promising approaches based on the HRV analysis [9]. Furthermore, the proposed method requires the use of PPG signal since it can be easily sampled from a wide range of sensors placed on the steering wheel. To evaluate the car driver's level of attention retrieved from the driver's PPG signal analysis, we designed a fully convolutional deep network that determines a saliency map of the driving scene. The results have highlighted the effectiveness of the proposed pipeline in assessing the driver's drowsiness level required by the driving scenario. Specifically, the system compares the level of attention reconstructed from driver's PPG signal with the level of attention properly calibrated to the driving scenario retrieved from saliency analysis alerting the driver if a risk occurs i.e. if there is a mismatch between the so detected attention levels.

2.9 Benchmarking of Computer Vision Algorithms for Driver Monitoring

As previously mentioned, a significant problem of developing DL pipelines in driving scenarios consists of designing suitable solutions for a real-time environment. DL techniques require a high workload that could hinder their application in the embedded automotive systems. However, there exist automotive -compliant devices designed to running demanding algorithms in automotive applications. In our research activity [93], we proposed a benchmarking evaluation of CV algorithms, which allowed us to perform a direct comparison between the performances of an automotive designed computational board (STA1295 Accordo5 MCUs) and a common laptop, intending to run CV algorithms on already existing hardware designed

for real-time applications in the automotive field. This work represents a first attempt to run such advanced algorithms in the board produced by STMicroelectronics Srl. By analyzing the processing times, we evaluate whether our automotive-compliant board is suitable to perform demanding algorithms or not. In the latter case, we aim to define arrangements needed to reduce the overall computation workloads and improve its ability to run more complex methods.

2.9.1 The Drowsiness Detection System

This section presents the pipeline developed to evaluate car driver's drowsiness. Our approach is based on assessing a driver's fatigue level through facial analysis by extracting facial component features (e.g., eyes, nose, mouth, etc.) from the human face image. Feature extraction is a key component of most CV applications, ranging from face recognition to age estimation [94], since providing a vast amount of information. In the automotive field, facial analysis systems have been widely used to measure the eye closure duration to find drowsiness hints. Indeed, previous research activities have demonstrated the existing correlation between vigilance and eye blink duration [95]. Inspired by the work of Soukupová and Čech [96], we computed the Eye Aspect Ratio (EAR) values, estimating the eye blinks, in order to assess the car driver's drowsiness. We collected video sequences of the car driver's face emulating a drowsy scenario at the first stage. To conduct our experiments, we used a common laptop and a development board based on an Accordo5 processor to compare their performances in terms of computing time. After recording the sequences, we reduced the video resolution from 1280×720 to standard VGA size of 640×480 pixels. The amount of memory allocated, related to the board, did not allow us to process videos with higher resolution. For this reason, we converted videos to a lower resolution. In order to detect facial landmarks, we cropped facial regions from video frames. Therefore, we load the OpenCV's Haar Cascade related to frontal face landmarks detection. Haar Cascade is an object detection algorithm based on the concept of features proposed by Viola and Jones [97]. Despite being an effective procedure to perform face detection in images and videos, Haar Cascade presents some drawbacks related to the demanding training time and the poor performances under difficult light conditions. We detected and extracted facial landmarks from frames of a video sequence from the dataset by using *Facemark*, the OpenCV's facial landmark API. Specifically, we used *FacemarkLBF* which represents an implementation of the algorithm proposed by Ren et al. [98]. The approach described in [98] consists of a set of local binary features and a locality principle for learning those features. The main novelty is related to learning discriminative local binary features to perform regression for the final output and elaborating facial landmarks independently. The main advantage of the proposed approach is the lower cost in terms of computation time since regressing local binary features is a very cheap operation. At this stage, we created an instance of *Facemark* class which is wrapped inside the OpenCV pointer, optimizing the memory management. We load a model related to landmarks detector trained on a considerable number of training images and the corresponding annotations. We repeat this procedure for each video frame of recorded sequences. We also estimate computing time in order to measure the performance of both devices. Although the Dlib library [3] is considered the "gold standard" to estimate facial landmarks in Computer Vision applications, it requires a high computational cost. For this reason, we adopted the *Facemark* API. Once detecting the 68-facial landmarks, we developed an effective drowsiness detection system

[3] Dlib library: `http://dlib.net/`

by considering the eye landmarks. According to [96], the Eye Aspect Ratio (EAR) value represents the distance used to determine if a person is blinking. Each eye is represented by 6 (x, y)-coordinates. Also, we choose the number of consecutive frames in which a driver closes his eyes. This threshold is not to be exceeded. When the driver closes his eyes, and the EAR value is below a given threshold for a number of consecutive frames, a warning message appears to alert that the fatigue of the car driver is high and could be dangerous. The EAR value is computed using the following equation:

$$EAR = \frac{\| p2 - p6 \| + \| p3 - p5 \|}{2 \| p1 - p4 \|} \tag{2.12}$$

where p1, . . .,p6 represent the 2D eye landmarks locations. The numerator indicates the Euclidean distance between vertical eye landmarks. The denominator represents the Euclidean distance between horizontal eye landmarks.

2.9.2 Resources and materials

Dataset

The dataset is composed of 12 videos. Specifically, this set contains video sequences of a face of people while driving. We acquired video sequences from drivers of different genders and ages. For completeness, we set up experiments also including drivers while wearing eyeglasses to evaluate the effectiveness of the drowsiness detection algorithm considering different scenarios. The facial camera was at a distance of approximately 30 cm from the subject. In particular, it was placed in front of the driver to record the subject's frontal face while driving. Each video is 100 seconds in length. The resolution is set to $720p$, and the frame rate is 30 fps. We emulate a drowsy scenario by performing eye blink closures while recording videos under high light conditions.

Hardware and devices

In this section, we briefly provide an overview of the used hardware, listing the main characteristics of each device. In order to evaluate the benchmarks, we used OpenCV (v. 3.4.3) installed on both systems, running C++ (std11) implementation of drowsiness detection algorithm to improve overall computational efficiency.

LAPTOP. Experiments were carried out on a laptop with an Intel Core i7 4710HQ CPU with 4 cores, 16GB of RAM and a N550JK motherboard, running Ubuntu 18.04 LTS.

ST BOARD. The other device involved in experiments is the development board STA1295 based on Accordo 5 processor produced by STMicroelectronics (software environment with embedded Linux) [88]. Specifically, the Accordo5 Evaluation Board (A5EVB) is a highly integrated "Car Radio," mounting the STA1295 version of the Accordo5 device, which is packaged in a 19 $\times$ 19 mm LFBGA Package with 529 balls, pitch 0.8 mm. The board is also composed of a 720p 10-inch Display Panel.

ACCORDO 5. The Accordo 5 multi-processor is a new line of digital-infotainment chips developed by STMicroelectronics [88]. The main advantage of the Accordo 5 chips is their integration of a dedicated, isolated ARM® Cortex®M3 core that secures the interface between the head unit and the main vehicle network. The microcontroller features built-in boot-code authentication, secure interconnection, and high-performance data encryption to manage secure CAN (Control Area Network) connectivity in real-time. Furthermore, Accordo 5 devices provide several attractive

Run	Video elaboration (PC)											
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**	**V11**	**V12**
1	24.6 s	25.5 s	26.7 s	24.1 s	25.9 s	31.8 s	27.3 s	28.6 s	27.1 s	27.0 s	27.1 s	28.1 s
2	24.4 s	24.8 s	25.5 s	23.4 s	26.0 s	31.1 s	27.8 s	28.4 s	27.3 s	27.6 s	27.3 s	28.2 s
3	25.1 s	25.797 s	25.5 s	23.3 s	26.1 s	31.6 s	27.7 s	28.6 s	27.1 s	27.0 s	27.5 s	28.2 s
Avg Time	**24.7 s**	**25.4 s**	**26.0 s**	**23.6 s**	**26.0 s**	**31.5 s**	**27.6 s**	**28.5 s**	**27.1 s**	**27.2 s**	**27.3 s**	**28.1 s**

Table 2.8: The processing time of a common laptop.

Run	Video elaboration (STA1295 Board)											
	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**	**V9**	**V10**	**V11**	**V12**
1	551.7 s	682.1 s	705.0 s	683.0 s	543.5 s	613.6 s	536.2 s	650.9 s	651.6 s	633.6 s	683.9 s	663.4 s
2	525.4 s	648.0 s	671.2 s	686.0 s	715.3 s	593.9 s	534.8 s	633.2 s	504.9 s	616.4 s	690.0 s	729.1 s
3	540.7 s	653.3 s	676.8 s	679.9 s	688.8 s	597.8 s	535.8 s	540.2 s	609.0 s	671.9 s	517.7 s	635.5 s
Avg Time	**539.3 s**	**661.1 s**	**684.3 s**	**682.9 s**	**649.2 s**	**601.8 s**	**535.6 s**	**608.1 s**	**588.5 s**	**640.6 s**	**630.5 s**	**676.0 s**

Table 2.9: The processing time of STA1295 Accordo5 embedded automotive platform.

features, such as smartphone mirroring (e.g., music and navigation services, etc.). In addition, the Accordo 5 family includes many state-of-art features, such as digital instrument clusters, consisting of complex and elegant displays which replace conventional dials and indicator lamps. Moreover, these chips integrate graphics, video, and audio functionality alongside Audio/Video/Navigation (AVN) head units and dual-screen capability to allow simultaneous user-interface plus rear-view camera with navigation and video previews, making driving safer and more comfortable.

YOCTO. The Operating System on STA1295 Accordo5 embedded automotive platform is an ad-hoc YOCTO Linux-based distribution (v. 1.8) [99]. Specifically, YOCTO is a collection of tools and meta-data that allows developers to build their custom distribution of Linux for their embedded platform. The main benefit of Yocto consists in building custom Linux OS for about any kind of computing device, allowing the creation of Linux images for all the major hardware architectures. In addition, it makes the creation of custom Linux distribution faster, easier and cheaper, which is a crucial aspect when developing embedded products. The main parts of the Yocto Project consist of the build system, the package meta-data, and the developer tools. Specifically, the build system uses a tool called *bitbake* to process the meta-data and produce a complete Linux distribution. It also builds the kernel, libraries, and programs that comprise a Linux distribution. In order to perform the deployment to the target device, it prepares the resulting software by placing it into appropriate bundles (including packages, images, or both). Moreover, Yocto provides application development and debugging to support developers.

2.9.3 Results and Discussion

In this section, we present and analyze the obtained results. In order to perform the proposed benchmark evaluation, we executed the described drowsiness detection algorithm 3 times, estimating the average time taken by both used devices to run it. Table 2.8 shows the computational time for each video of the dataset. The average time is highlighted in bold. As evident, the results revealed that the performance of

the common PC is quite efficient. We observed that the maximum running time required to compute drowsiness detection algorithms is 31,515 seconds in relation to the video in which we performed the highest number of eye blink closures (Video 6), while the minimum running time is about 24 seconds. On the other hand, our experiments reported poor performances concerning the embedded STA1295 board. As expected, our tests show that the automotive-designed board yields a longer computational time than the common laptop. In general, running a Computer Vision algorithm on our embedded system faced several limitations in terms of memory and power consumption since it was not developed to execute demanding algorithms. In analyzing the performances of the used board, it turned out that the most demanding part was related to I/O operations. To minimize the computational costs, we optimized I/O operations to increase the processing efficiency. In this respect, we performed the following steps:

- video decoding, computing CV algorithms, frame pushing on a buffer;
- closing reading stream
- frame popping from buffer and writing on file system

Despite attempts to increase the overall performance, the automotive-grade board takes a long time to compute the implemented drowsiness detection algorithm. As seen in Table 2.9, the results reported an average running time of over 530 seconds for each video which is not acceptable for a real-time application. Indeed, the resulting performance results from multiple factors, spanning over OpenCV implementation and video properties (resolution, fps, etc.). Although it is feasible to cross-compile the OpenCV source code on embedded devices, memory constraints and other architectural considerations may pose a problem, complicating optimizing CV functions on a new target device. Furthermore, video properties, such as resolution, fps, etc., will directly affect the efficiency of the overall system. As previously reported, we reduced the resolution of videos (Standard VGA resolution) since our automotive-grade device presents limited resources. To sum up, we used a ready-to-go development board taking advantage of its integrated features to execute a Computer Vision algorithm for evaluating the drowsiness of a car driver. Despite its limitations and the poor results reported in Table 2.9, our findings suggest that this device could be employed for automotive applications based on running CV algorithms. In fact, by developing a fierce trade-off with respect to memory usage, I/O bandwidth, and algorithm complexity, we will provide a significant improvement to maximize the potential of Computer Vision functions on our embedded platform.

2.10 Future Works

There are a few directions for further studies. Specifically, we aim to implement more advanced CV algorithms to provide a complete study regarding driver status analysis. Moreover, considering the remarkable success of DL architecture over the last years, our future works will focus on optimizing the use of DL approaches on our automotive board based on Accordo 5 processor [88], as well as reducing the computational cost for the development of a valuable tool for the automotive industry. Moreover, considering that most artificial intelligence algorithms require real-time behavior, with latencies on response times on the order of a few milliseconds. One of the future goals is to dedicate future research to ensure a low latency, which is fundamental for systems with real-time behavior, making the embedded system

more autonomous, with higher decision-making. In the context of visual saliency, more investigations are underway to improve the saliency analysis by extending our dataset, including further driving scenarios in different domains.

2.11 Conclusions

In this chapter, we detailed the development of our proposed pipeline to determine driver drowsiness. The implemented solutions have exploited the main techniques of CV and DL to acquire visual information and process them. The main advantage of the proposed solutions is the use of non-intrusive and affordable tools for acquiring physiological signals, such as PPG. As reported, devices to extract driver physiological conditions must undergo several requirements to facilitate their embedding into the vehicle environment. In a broad perspective, the acquisition devices consisted of automotive-compliant sensors placed on specific vehicle points to capture physiological information. The most suitable places are the steering wheel, where the driver maintains his hands to keep control of the vehicle, and the passenger seat. However, the disadvantage of installing sensors into a vehicle is linked to inaccurate data acquisition. For example, the driver may not maintain both hands on the steering wheel during driving. This issue does not allow us to collect physiological properly, leading to incomplete data extraction. Sensors located in the driver's seat also fail to capture adequate information since the driver may be assuming non-compliant positions for data acquisition. These considerations have led to the development of advanced solutions based on DL techniques to analyze a driver's psychophysical state. In particular, we based our work on the study of Wu et al. [68] which is related to Video Magnification for the mapping of facial features to reconstruct the PPG signal. Thus demonstrating that it is possible to acquire information about a user's state from visual data. Moreover, such techniques encourage the use of non-expensive tools, such as a low-resolution camera, to extract data about the driver's face. Another significant advantage of the proposed methods is the adoption of automotive-compliant tools devoted to running ADAS functions. As introduced, AI has achieved high precision and accuracy in processing important information in images and video data. However, the high accuracy comes at the price of high computational costs. As a result, dedicated hardware devices, from the application of specific processors, are needed to optimize complex workloads of the AI methods. In the automotive industry, the current capabilities of automotive-grade hardware devices are limited. For this reason, the scientific community is working intensely to overcome the constraints in intelligent in-vehicle technologies. We designed the proposed solutions by taking into account the installation of such functions in the car environment. In this regard, we used a development board, based on Accordo 5 processor [88], that includes integrated features for running CV algorithms. Specifically, this board allows us to have ready-to-use hardware, not requiring any design activity, hardware development, or testing. In addition, it assures high flexibility: the same hardware platform can adapt to different types of applications without requiring hardware changes.

In this chapter, we have detailed most of our research activities, attempting to improve the state-of-the-art in the automotive field. However, note that some of the works listed in the publications section are not reported here since they present pipelines similar to those of the mentioned works.

Chapter 3

Medical Imaging

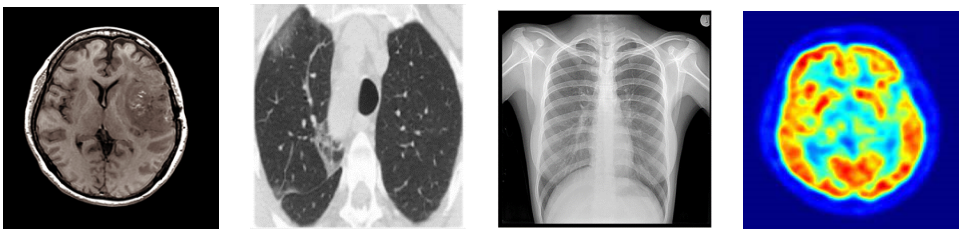
3.1 Overview

Since 2012, DL has gained impressive results in several fields ranging from bioinformatics to medical [100]. The spread of DL techniques has led to the development of algorithms with a performance similar to or even better than human operators one. Their success is largely due to their strong ability to automatically learn feature representation from data in order to perform complex tasks, such as image classification, localization, detection, and segmentation. In addition, DL algorithms process massive quantities of data in a reasonable amount of time. Hence, it is evident how DL methods could be advantageous in the clinical field, especially at providing the rapid diagnosis of patients. Furthermore, the application of DL algorithms has allowed solving challenging tasks (e.g., disease detection in medical images), providing a significant contribution in the area of Medical Imaging. Although Medical Imaging has recently received great interest from the scientific community, the first applications of AI algorithms in this field can be traced back to 60s [101], where a ML approach was used. Since then, a wide range of notable solutions exploiting the DL frameworks[1] have been proposed, especially in the 90s [102], [103]. At the time, scientists could not benefit from advanced computational resources such as the contemporary ones, thus explaining the initial skepticism towards DL paradigm. In fact, a more fervent research activity have been conducted in medical field only recently.

Medical imaging refers to a variety of different techniques used to photograph the internal structures of the body in order to facilitate the diagnosis, the monitoring and the treatment of health conditions. Medical Imaging, which includes radiography, ultrasound, and Magnetic Resonance Imaging (MRI), has always required the flexibility of human expertise to detect biological abnormalities, as the medical images often present a noisy backgrounds and quality issues such as specular reflections[2]. Indeed, it is a time-consuming and difficult process. In this regard, DL-based algorithms may automate the image analysis in a more reliable way. In fact, DL is profoundly changing the medical field. There are numerous advantages to be derived from its application to medicine: from increased productivity, to greater diagnostic accuracy, to the chance of facilitating access to diagnostic tests (even in places and for people who cannot benefit from them due to geographical, political and economic barriers). Despite achieving remarkable results, DL has encountered some obstacles in processing medical data due to their peculiarities, which disturb the image interpretation, hindering its diffusion. In fact, the characteristics of medical images could affect the performance of standard DL algorithms to accurately

[1] These techniques are related to DL architectures consisting of two or three layers, which are mostly not considered "deep" model nowadays.

[2] This is a major problem in medical field, where reflections from surface lead to an image quality reduction.



(a) (b) (c) (d)

Figure 3.1: Examples of medical imaging modalities. (a) MRI, (b) CT-scan, (c) X-Ray, (d) PET.

locate an object or region of interest, especially for identifying abnormalities in an unstructured scenario.

In this dissertation, we cover the main areas of research in the medical field. Firstly, we target the problem of data *interpretability* in order to explain DL model decisions to physicians. In particular, we discuss the importance of providing precise results to clinicians, who often are reluctant towards DL techniques. Their skepticism relates to DL methods in finding correlations between data and processing because they often present an ambiguous nature. This problem is called the "black box" problem. Hence, understanding how DL approaches to generate a given output by processing input data has become pivotal in the scientific community in order to develop effective solutions for data interpretability. Secondly, we addressed the problem of classifying multi-modal images created by means of tools with different characteristics. In fact, classifying images generated by different tools can result in inaccurate algorithm performance. After describing the main traits of medical images in detail, finally, we discuss how the application of DL algorithms is being inhibited mainly due to the non-availability of high-quality labeled datasets, addressing the challenges that specialists had to deal with.

3.1.1 Imaging Modalities

Medical imaging can be defined as a technique for developing visual representations of the human body's interior to diagnose health problems and consequently monitor how the patients respond to a certain kind of treatment, allowing the clinicians to provide diagnosis and clinical treatments to patients without dangerous side effects. Medical imaging encompasses different imaging modalities that operate differently in order to view specific parts of the human body. In Fig. 3.1, we reported examples of existing medical imaging modalities.

Common imaging modalities include:

- Computed Tomography (CT)
- Magnetic Resonance Imaging (MRI)
- X-Ray
- Positron-emission tomography (PET)
- Ultrasound

Computed tomography (CT) is a diagnostic imaging technique that allows us to examine every part of the body (brain, lung, liver, pancreas, kidneys, uterus, arteries

and veins, among others) for the diagnosis and study of tumors and many other diseases. It is an X-Ray examination in which a computer processes the data collected from the passage of several beams of X-Rays in the affected area, reconstructing a three-dimensional image of the different tissue types.

Magnetic Resonance Imaging (MRI) is used to produce high-definition images of the interior of the human body, focusing on soft tissue [104]. The introduction of MRI imaging into clinical practice has profoundly changed and expanded neurological diagnostics. It represents a harmless method because it uses magnetic fields without ionizing radiation and exploits the physical properties of the hydrogen atom subjected to magnetic fields and radio-frequency pulses. Unlike CT scans, MRI can be performed with or without intravenous injection of a contrast agent which has no or few side effects. In addition, contrast medium facilitates visualization of inflammatory processes and highly vascularized tissues, such as tumors. MRI is currently indicated as the "gold standard" among diagnostic methods of studying a wide range of brain pathologies due to its intrinsic characteristics, non-invasiveness, and diagnostic sensitivity. However, as will be shown below, there are some further considerations on this point. The Diffusion Tensor Imaging (or DTI) and Functional Magnetic Resonance Imaging (or fMRI) subgroups are part of the MRI family. DTI is an MRI technique that allows the analysis of diffusive properties and directionality of flow of water molecules within tissues in vivo, and is presented as an important tool for studying the microstructural architecture of brain structures in both physiological and pathological conditions. fMRI consists of visualizing the hemodynamic response (changes in the oxygen content of the parenchyma and capillaries) related to neuronal activity in the brain. The change in the oxygenation state of hemoglobin in red blood cells is the theoretical principle of the BOLD (Blood Oxygen Level Dependent) effect, on which fMRI is based and which is used as an endogenous contrast agent.

X-Ray is an investigation that uses the properties of a particular type of ionizing radiation, X-Rays, to make an impression on a film (plate) in order to obtain images of the human body depicting both bones and specific organs. In the X-Ray, the image obtained is negative: the parts of the body which are consistently denser (e.g. bones) appear clear. On the contrary, the soft tissues appear gray, and the organs that are crossed by X-Rays (e.g. lungs) appear dark. Nowadays, digital X-Rays are increasingly used. An X-Ray image is processed by a computer that views and then stores it in digital form.

Positron-emission tomography (PET) is a method of diagnostic imaging for the early detection of tumors and evaluation of their size and location. The examination is based on the administration of radiopharmaceuticals, characterized by the emission of particles called positrons. PET involves administering a small amount of a radioactive substance (radiopharmaceutical) to investigate the functional characteristics of organs and apparatuses in which the radiopharmaceutical is localized. After being administered intravenously, the radiopharmaceutical is distributed throughout the patient's body, allowing to obtain diagnostic images analyzed by clinicians.

Ultrasound uses acoustic waves to reconstruct images of the inside of a body. A beam of high-frequency acoustic waves is directed into the interior of the body. The waves reflected from the different tissues are then picked up by a transducer placed on the body's surface and transformed into images. The advantages of visualization by ultrasound are the minimal cost of the equipment compared to other machines and the possibility of obtaining images in real-time without using any invasive tools.

3.1.2 Medical imaging challenges

DL decision support systems are garnering great interest due to their high diagnostic accuracy in specific clinical contexts. However, some potential drawbacks may critically affect the hypothetical advantages of applying AI systems in the medical domain. In light of the current lack of studies on the side effects of applying these new decision aids in medical practice, this dissertation summarizes the main unexpected consequences that could be derived from their extended application. These consequences are especially related to predictive DL models in which high accuracy is often inversely proportional to the transparency of the path leading to the predictions. The main unintended consequences related to the use of DL systems vary from the inherent uncertainty of the data, used to train these systems, to the inadequate interpretability of their responses, as well as to the risk of an excessive tendency not to rely on such systems. In this regard, we explain the critical elements that need to be addressed by the Medical Imaging community, deepening the knowledge about the three major challenges encountered during our research activity. More broadly, the critical issues are related to:

Data Interpretability. The ability to determine how DL models generate their outcomes is known as "interpretability." The AI research community has recognized the problem of *interpretability*, thus feeding an increasingly flourishing area of investigation. Researchers/Users should have to understand how a model draws its conclusions without fully recourse to their experience or monitoring the dataset to compensate for limited interpretability. Despite yielding good results, DL models operate as a "black-box" as their process to make predictions on unseen data leads to an unclear procedure. In this regard, researchers often struggle to establish why an algorithm produces a given answer. Interpretability is a crucial property for ensuring that the predictions are unbiased. In fact, interpretability is a reliable tool for detecting bias in DL models, preventing discrimination against underrepresented data classes. Furthermore, interpretability is a valuable measurement of the effect of trade-offs in a model. Additionally, as we will discuss in the following section, there is a strong correlation between interpretability and robustness. In the medical field, these two variables are necessary to ensure that DL techniques can be applied in various contexts. Over the years, several techniques have been proposed for enhancing the degree of interpretability in DL models. In Section 3.2, we discuss the most common of these techniques.

Multi-modal data. With the continuous growth of medical imaging technologies, a large collection of medical data has been produced with rapid incidence. Medical images present multiple modalities according to the method used to generate them. As previously introduced, common image modalities are ultrasounds, X-rays, and MRI. Since the medical acquisition techniques are numerous and performed using devices of different manufacturers under non-standard settings, clinical repositories collect a massive amount of heterogeneous medical data. Despite the advances in technology, the involvement of a human operator to perform image analysis is still predominant in the medical domain, leading to a time-consuming and error-prone process. In this regard, the application of DL algorithms may automate the ability to filter images from a vast collection of data in order to support clinician's diagnosis for future queries. However, images deriving from multiple inputs present several traits, ranging from the resolution to the number of the slice, which may inhibit the suitability and nature of DL solutions. In this regard, we designed an effective DL pipeline to perform *modality classification*. In Section 3.4, we address the difficulties experienced in processing multi-modal data and show some

preliminary results.

Non-availability of labeled dataset. A vast majority of AI algorithms applied in the medical field are based on the supervised learning approach. This kind of approach implies that data used to train models are associated with a *ground truth*. By definition, the term ground truth is used to indicate information that is empirically true. In the context of medical imaging, ground truths (or labels) are annotations defined by clinicians after they performed direct observation of specific examinations (e.g., radiography). Nowadays, the AI community must cope with the scarcity of clinically labeled datasets, which has reduced the spread of DL approaches in the clinical environment [105]. Generally, medical experts performed image-labeling through a manual process, relying on their experience and advanced knowledge of the data. Manual labeling is currently the commonly used approach in acquiring labeled imaging data for AI applications, as researchers observed that manual labeling produces reliable labels. However, this human intervention is time-consuming due to the large dataset that needs to be annotated. On the other hand, semi or fully automated data-labeling can reduce time and costs concerning expert involvement, but it produces sparse and noisy annotations. In addition, different tasks require different forms of annotation that often require human expertise. For this reason, the establishment of gold standards for image labeling remains an open issue. In Section 3.5, we describe in detail the critical aspects we faced during the development of DL pipelines for medical applications, and we outline how we dealt with the lack of adequate datasets.

3.2 Interpretability

3.2.1 Defining Interpretability

The ability to "think" is a crucial facet of human intelligence, which permits us to learn from experience, make predictions, and rationalize in order to find an explanation when unexpected events occur [106], [107]. Thinking is the main strength of the human brain because it provides a logical explanation for a given output throughout understandable steps. In the scientific field, the counterpart of the human brain is the DL model. The ability to understand what a DL model learns is indicated as *interpretability*. Model interpretability has become essential to explain how features are involved in modeling and how they affect the prediction of a neural network. Indeed, an exhaustive explanation supports the comprehension of the model's predictive decisions and increases user trust in these deep models. Lipton et al. in [108] attempt to provide a formal meaning for interpretability. They posed that interpretability incorporates two main concepts: *model transparency* and *model functionality*.

- *Model Transparency* connotes the capability of understanding the mechanism behind a DL model, it is defined in relation to the following properties:
  - *Simulatability*, which denotes the act of reproducing every calculation step that leads to a prediction by using both the input data and the model. This property is fundamental for having a comprehensive knowledge of the changes behind a DL process.
  - *Decomposability*, which indicates whether an intuitive description is provided for each component of the model, such as input, parameter, and calculation.

  - *Algorithmic Transparency*, which denotes the fundamental property of transparency. It details the ability to explain factors that affect the learning process of an algorithm.

- *Model Functionality*, it is based on three aspects:

  - *Text Explanations*, which refers to an approach to explain the model predictions in natural language, providing a justification for them.
  - *Visualization*, which provides the visualization of the model parameters to help us better understand the working of a model, enabling us to identify what is important in the decision-making process.
  - *Local Explanation*, similar to the visualization, focuses on the change to inputs and outputs, observing local variations instead of explaining the entire mapping of a model.

In the medical domain, the *interpretability* is a crucial property for the development of reliable DL models [109]. The current literature has highlighted the motivations to prompt the design of interpretable models, especially in a challenging field, such as the medical one. Generally, interpretability may provide valuable insights to data used for training a DL model and ensure the quality of protocols related to AI systems. The latter is a critical point in medical domains due to the easiness of fooling DL algorithms, for instance, by changing a pixel in an image producing a system error (*adversarial attack*) [110]. Due to the fact that clinicians often have reservations about AI models as they are not able to understand their processes and how they determine outcomes, interpretability induces a more increasing trust in deep models [108], [111]. As the application of these technologies grows in clinical routines, several approaches have been proposed over the years, ranging from image segmentation to monitoring disease progression. Interpretable approaches provide helpful information for assisting clinical professionals as interpretability confer trustworthiness, thus promoting effective adoption in practice [112], [113].

3.2.2 State of the Art

A consistent amount of studies based on the interpretability has tackled the problem of understanding how a DL network makes predictions and why it produces them [111] [114] . In other words, researchers addressed the dimension of *model transparency*, which is indicated as the most significant property. The deeper the model, the lower the *algorithmic transparency* [115]. These studies have also shown that there is a close correlation between the interpretability of a model and its robustness. Specifically, the scientific community has highlighted that the robustness of a model implies a greater interpretability. Taking into account that DL models show vulnerabilities to perturbed data that may compromise their effectiveness [116] and trustworthiness, a large amount of research works investigated the adversarial robustness, demonstrating that adversarial defense strategies, such as adversarial training, could lead to a better understanding of the model strategy. Motivated by these findings, we report recent studies that outline the improvement of the DL interpretability. In addition, we provide a discussion on the relationship between adversarial robustness and interpretability. With regard to the model interpretability, Zeiler and Fegus [117] proposed a novel visualization method in order to interpret the feature activity in intermediate layers. Specifically, they adopted a Deconvolutional Network (deconvnet) to map the features to the pixel space leading to a given

activation in the feature maps. This work [117] proved that the improvement of the model transparency permit us to build better models. In [118], the authors adopted a method to provide an interpretable testbed designed to revealing the existence of interpretable cells in LSTM models. Specifically, the authors demonstrated that some cells learned interpretable features, contrary to other units, which produced less easily interpretable outputs. They used a method to visualize the activation of individual units after performing the training of the LSTM model, which was fed using one character at a time on different texts. Another work devoted to providing features visualization is [119], where the authors proposed two tools to accomplish this task. In particular, the first tool produces activations for each layer of a trained DNN which was fed using images or video, whereas the second tool performs better visualization of the learned features, showing how DNNs process the input data. The results suggested that the combined effect of these tools could promote a greater understanding of DNNs computing. Simonyan et al. [120] used two techniques for visualizing features of image classification models (ConvNets). The first method creates an artificial image sample related to a specific class. The second technique provides the saliency map of the image, discriminating what features represent the given class. Finally, the *GraphCut* color segmentation is generated by using both image and the corresponding class saliency map. The color segmentation helps us to visualize only the most discriminative part of the object depicted in the image.

Considering that clinicians find hard to understand and trust on DL complex models due to the lack of intuition and explanation of their predictions [121]. One of the most used technique in healthcare for providing local interpretability is Local Interpretable Model-Agnostic Explanations (LIME) [111]. The intuition behind LIME algorithm is that global behaviour of a DL model is complex to understand, while the local behaviour is much easier. LIME is a recent technique that explains the model's prediction by introducing perturbed data in order to understand how prediction are affected by changes. Specifically, LIME establishes the contribution of each feature to the final prediction. This indicate what feature has the most impact on the model's output, providing local interpretability. Despite being effective, LIME algorithm is computationally prohibitive and difficult for the explanation of the entire dataset during the training process. In order to overcome the spurious learning of continuous data, Ross et al. [122], getting inspired by LIME approach, proposed the setting constraints for the input gradients, which are faster than sample-based methods, such as LIME. The proposed method used binary masks that determine whether an input feature is irrelevant to the classification of that example, according to the annotation of a human expert. Furthermore, the authors designed an automatic method that uses different masks for generating models having different decision boundaries.

In [123], the authors attempt to provide an explanation of the model behavior. Specifically, they assess the importance of data points from training data finding what data points contribute most to classification errors. They also introduce imperceptible perturbations on images (adversarial examples) revealing insights about how model's prediction would be affected if data were altered, and how it extracts the training data. With regard to adversarial examples, an extensive amount of literature suggest that the model interpretability is closely related to robustness, i.e., robust Deep Neural Networks (DNNs) tend to have more interpretable gradients [17]. In this regard, Anil et al. [124] argued that performing the model training process by using Lipschitz constraints ensures its robustness. The authors Ross et al. [125] demonstrated that gradient regularization increases the model robustness as well as

its interpretability and its ability to learn features which are better aligned with human perception. A flourishing research area has explored the correlation between adversarial robustness and interpretability. As introduced, DL networks present some vulnerabilities because they are sensitive to the learning techniques and data used. For example, the addition of new data can significantly change their predictions. Specifically, perturbed data may introduce unexpected critical issues which lead to poor results. In this regard, many solutions have been proposed to tackle the problem of adversarial attacks, defining advanced techniques which contrast the effect of noisy data during the training. In this instance, the adversarial training has been proposed as an effective solution to overcome the problem of adversarial attacks. Kuranin et al. [126] introduced adversarial examples during the training process in order to improve the model robustness. Chan et al. [127] designed an effective approach, Jacobian Adversarially Regularized Networks (JARN), devoted to improve the model robustness by adversarially regularizing the Jacobian of the network. The procedure permits to increase the model robustness to Projected Gradient Descent (PGD) attacks. In addition, recent works have also observed that regularizing the Frobenius norm of the Jacobian matrix affects the overall generalization error [128].

3.2.3 Adversarial attacks

Despite being the leading tool in the general imaging and Computer Vision domains [129], [130], DL models present a major weakness related to their vulnerability towards adversarial data. Adversarial attacks can fool Deep Neural Networks (DNNs) by introducing small perturbations during the input stage to produce poor results in the testing/validation stage. In Fig. 3.2, we depicted the effect of an adversarial attack on images. In this regard, the concept of *robustness* has become a key point in the context of safety-critical environments and applications where the adversarial attack may lead to significant risks.

Formally, let $x_0 \in R^d$ be a data point that belongs to class C_i. Define also a target class C_j. An *adversarial attack* is defined as a mapping function $A : R^d \rightarrow R^d$ such that the perturbed data

$$x = A(x_0) \tag{3.1}$$

is misclassified as a data point of class C_j.
Specifically, adversarial attacks are divided into two main categories:

- ***White box attacks.*** It generates adversarial examples by taking advantage of its full knowledge about the parameters and architecture of the model. Within scientific literature there exists a considerable amount of work that suggests the effectiveness of two adversarial attacks that fall into the category of white-box attacks: the so-called *Fast Gradient Method (FGSM)* [116] and the *Projected Gradient Descent (PGD)* [16].

- ***Black box attacks.*** In contrast to white-box attacks, a black-box attack has no access to DL models during the training phase and has to deal with a limited knowledge of the model [16]. An attack with black-box constraints is often modeled around querying the model's inputs, observing the labels or confidence scores.

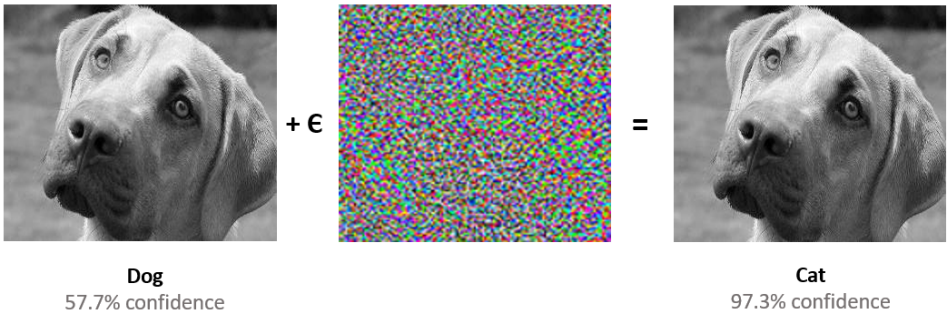


Figure 3.2: Examples of adversarial attack. An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a dog as a cat

3.2.4 Defense Techniques

The scientific community has pointed out that robustness to adversarial attack implies robustness in the decision, which in turn leads to interpretable feature maps, according to *explainable AI* theory [125], [131], [132]. In this instance, we address the lack of robustness of DL models by using an effective strategy, known as *adversarial training*, which is currently one of the most effective defense technique to hinder the effect of adversarial attacks [133]–[135]. Specifically, adversarial training adds perturbed examples to the training set and then retrains the model to enhance robustness, allowing the model to effectively cope with the adversarial attacks. The idea behind the adversarial training is to solve the so-called *min-max problem* [136]

Consequently, recent studies investigated adversarial examples to provide countermeasures for adversarial attacks and develop a well-defined DNNs model. Furthermore, recent literature has deepened the analysis of the visual interpreting model's gradients [137], demonstrating that adversarial training generates gradients of the output, with regard to the input, which lie closer to the image manifold, forcing model decisions to align better with input patterns. The principal defense techniques are:

- ***Distillation.*** Distillation is an adversarial training procedure where one model is trained to predict the output of probabilities, thus extracting class knowledge from these probability vectors, to transfer it into a different DNN architecture during training [138], [139].
- ***Adversarial Training.*** Adversarial Training is the most applied defense mechanism against adversarial samples. This technique is devoted to improving the model robustness by introducing adversarial images during the training step [16], [140].

Several studies have proved that adversarial trained DNNs learn feature representations that are better aligned with human perception. In the following section, we tackle adversarial attacks by applying adversarial training to improve the model robustness, which, as stated previously, affects interpretability: input gradients from adversarial trained DNNs appear to be more interpretable than their standard counterparts [17], [125], [137]. Moreover, we estimated the model interpretability through evaluating its robustness by regularizing the Frobenius norm of the Jacobian matrix, which influences the overall generalization error [128].

3.3 Enhancing model robustness

The work herein described addresses the identification of oncogene-addicted tumor lesions from lung CT scans using a DL approach as well as the improvement of model robustness by using the adversarial training technique [141]. This study results from a collaboration with Professor Concetto Spampinato, head of the *PerceiveLab* group of the University of Catania. In this regard, our contribution is limited to targeting the problem of model interpretability, whereas the *PerceiveLab* team designed DL models to extract cancer lesions to be classified either as gene-addicted or not gene-addicted. Specifically, we handled model interpretability by designing a defense technique that contrasts the vulnerability of DNNs to adversarial attacks. In this regard, our findings confirmed that adversarial robustness entails feature interpretability, as suggested by explainable AI theory [17]. The approach is thus able to genomically characterize lung cancers while providing interpretable decisions.

3.3.1 Deep Learning models

In this section, we briefly described the overall pipeline used to discriminate the gene-addicted lesions from non-gene-addicted ones. Since this section is intended to providing an overview of model interpretability, we did not report all details regarding the proposed DL architecture but rather confirming the efficiency of adversarial training at improving model robustness.
The segmentation model. The segmentation network is devoted to extracting lesions from a CT scan. The segmentation model is based on the Tiramisu network [142], i.e., a 2D fully-convolutional DenseNet [143] based on the U-Net architecture [144], consisting of a downsampling path for feature extraction and an upsampling path for output generation, with skip connections that help to preserve high-resolution details by reusing feature maps between the two paths. Unlike the standard Tiramisu architecture, the segmentation model includes:

- 3D (rather than 2D) convolutional layers to process the whole CT sequence;
- Residual squeeze-and-excitation layers [145] used to emphasize relevant features and improve the representational power of the model;
- deep supervision [146] to the upsampling layers to deal with the vanishing gradient problem.

The classification model. The lesion classifier is a multi-scale 3D CNN that receives a crop of the CT scan corresponding to a lesion identified by the upstream segmentation model and it predicts whether the tumor is gene-addicted or not. Each branch of the model consists of a cascade of 3D depthwise separable convolutional layers with ReLU non-linearities, sharing the number of feature maps but differing by kernel size (3, 7 and 11). Padding is added to ensure that feature map sizes are consistent across the three branches. After multi-scale analysis, the output feature maps are summed and rescaled to a volume of size 4×4×4 through adaptive max pooling. A final fully connected layer maps the pooled features to a vector of size 2, with gene-addicted and non-gene-addicted class scores.

Dataset. The CT dataset is composed of 73 CT scans of lung cancer patients, diagnosed by biopsy. For each biopsy, the presence of a specific genomics profile is noted, with 21 out of 73 lesions defined as oncogene-addicted, i.e., characterized by a single dominant driver gene (either EGFR or ALK or ROS-1 or BRAF). The

Model	Accuracy (%) ↑	$\|\|\mathcal{J}(\mathbf{x})\|\|_{\mathbf{F}}$ ↓
No segm	58.00 ± 4.47	5.60 ± 4.27
Bottleneck	66.00 ± 5.47	11.06 ± 3.75
Ours	82.00 ± 8.36	0.58 ± 0.37
Ours + adv. training	78.00 ± 4.47	0.35 ± 0.36

Table 3.1: Classification results. Mean classification accuracy (in percentage) and Frobenius norm of the Jacobian matrix.

remaining 52 cases are not oncogene-addicted. An expert oncologist annotated each CT scan by drawing the contours of the biopsied lesion and selecting a subset of slices containing it.

More details related to both models and dataset can be found on [141].

3.3.2 Dealing with adversarial attacks: the proposed strategy

In this work, our efforts have been spent to improve the base classification model by enforcing robustness to adversarial attacks and interpretability. As previously mentioned, these two aspects are necessary to enhance trust in an AI model, especially in the medical domain. To accomplish this, we employed a white-box attack mechanism – PGD [16] – based on adversarial training [147] that aims at generating adversarial examples, through input perturbations, to make the model misclassify lesions during training.

Formally, let x_t be the perturbed input data after t iterations of the algorithm, starting from the original input $x = x_0$. Each PGD iteration adds a perturbation to the data at previous iteration and it projects the new data to a point within a L_2 hypersphere with radius ε around x_0 to ensure similarity to the original input. The iterative perturbation process can be described as:

$$x_{t+1} = \Pi_{x_0,\varepsilon}(x_t + \varepsilon \operatorname{sign}(\mathcal{L}_{\mathrm{XE}}(x_t, y)) \tag{3.2}$$

where t is the iteration number, y the label of the input sample, and $\mathcal{L}_{\mathrm{XE}}(x_t, y)$ the binary cross-entropy loss of the lesion classifier. $\Pi_{x_0,\varepsilon}(\hat{x})$ is defined as:

$$\Pi_{x_0,\varepsilon}(\hat{x}) = \begin{cases} x_0 + \frac{\hat{x}-x_0}{||\hat{x}-x_0||_2}\varepsilon & \text{if } ||\hat{x} - x_0||_2 > \varepsilon \\ \hat{x} & \text{otherwise} \end{cases} \tag{3.3}$$

During training, we feed the PGD-generated samples to the model with the original y label. This procedure forces the model to deal with targeted attacks at training time and to learn features that are less susceptible to input perturbations, resulting in increased generalization, robustness, and interpretability.

Training procedure. During training, the segmentation model receives a tensor of size $256 \times 256 \times 30$, obtained by resizing the height (sagittal axis) and width (frontal axis) of the original CT scan and using the 30 slices (along the longitudinal axis) around the annotated lesion as provided by the radiologists. The classification model is trained on the individual lesions identified by the segmentation model, padded by 5 voxels (to account for under-segmentation), and resized to $64 \times 64 \times 30$. More details can be found on [141]. Finally, perturbed samples for adversarial attacks are generated by setting $\epsilon = 0.1$ in the PGD algorithm.

Lesion classification results. We then evaluate the accuracy of the proposed lesion classifier by using the following baselines:

- *Classification without segmentation*, i.e., the multi-scale lesion classification model, operating on the entire CT scan.
- *Bottleneck features from the segmentation model*, i.e., forwarding the bottleneck features (size $24 \times 3 \times 16 \times 16 = 18{,}432$) of the segmentation model to three fully connected layers of size 1024, 1024 and 2 for lesion classification.

3.3.3 Evaluating robustness and interpretability

To measure the robustness and interpretability, we evaluated the average Frobenius norm of the Jacobian matrix $||\mathcal{J}(x)||_F$, which estimates how the model is affected by input perturbations. Basically, the smaller the values of the matrix, the less added perturbations affect the output. In a nutshell, the Frobenius norm consists of squaring all of the elements in the Jacobian matrix, taking the sum, and computing the square root of this sum. This computation describes the the L^2 penalty of the concatenated gradient vectors [148]. Results are reported in Tab. 3.1 and show that our approach outperforms both baselines in terms of accuracy and robustness. We can notice that using the segmented lesion leads (last two lines in Tab. 3.1) to better robustness and interpretability (as demonstrated by the achieved Jacobian norm values). This can be explained by the fact that using lesion-masked information during classification prevents the representations learned by the classifier from being inconsistent with lesions, which is one of the main reasons for scarce generalization and overfitting in these applications [149]. Additionally, adversarially training our classifier yields even lower Jacobian norm values at the expense of classification accuracy (82% vs. 78%). The accuracy decrease is expected since adversarial training forces the model to avoid using non-generalizable features, but it enhances generalization capabilities; thus, it is preferred.

3.4 Multi-modality classification

3.4.1 Overview

Modality classification consists of identifying the type of medical images by extracting a set of discriminant visual features. In the clinical field, a vast amount of medical images is producing annually, generating huge hospital repositories with a large collection of valuable information. With the continued growth of these image repositories, the development of automatic tools for grouping medical images according to specific characteristics has become an urgent need. More broadly, the role that these data are taking in clinical decision making and treatments assessment has prompted the designing of advances approaches for performing image retrieving and classification in the medical imaging domain. Traditional methods are based on the selection of hand-made features and require prior domain knowledge. More recently, the growing interest in DL technologies has led to the development of several approaches to accomplish these tasks in a much more efficient and data-driven way. This section presents some preliminary results related to the research activity conducted under the supervision of Dr. Daniele Ravì, Senior Lecturer at the University of Hertfordshire (UK).

3.4.2 State of the Art

Several advanced solutions have been proposed over the years regarding the development of effective classification systems for the modality classification of medical images [150]. The main benefit of modality classification is the ability to accurately identify data regarding a specific modality with significant time-saving benefits.

In the literature, there exist two main approaches for modality classification based on: (i) hand-crafted feature-based approaches and (ii) DL approaches.

Approaches based on hand-crafted features. Recently, several researchers have developed efficient methods for modality classification in order to segregate images concerning a specific modality on benchmark datasets [150]–[152]. In this regard, the authors of [153] implemented a modality classification system consisting of model classifiers: one for grey-scale images and the other for color images. Despite achieving remarkable results on *CISMeF* database, the model often misclassifies MRI and CT scan classes as suggested by the results of the confusion matrix. Another drawback of the [153] is related to the expensive workload of the model classifiers. Another method for image modality classification was proposed by [154]. The authors defined a framework using SVM and KNN classifiers in combination with a fuzzy rule-based technique. In this work, the authors performed image modality classification considering five types of medical data: CT-scan, X-ray, Ultrasound, MRI, and Microscopic images.

Han et al. [155] proposed a modality classification method by using both visual features and textual features, such as the binary histogram of some predefined vocabulary words from image captions. The combination of both input data based on the refinement procedure has led to better results than others.

Approaches based on Deep Learning. Chiang et al. [156] proposed modality classification technique using CNNs. In this regard, they collected images with different modalities, including CT and MRI. The results reported that the CNN model achieved good results, confirming its effectiveness.

Van et al. [157] investigated the Axial CNN for performing cross-modality learning. The used architecture consists of an autoencoder that extracts a standard representation for data coming from multiple modalities. To conduct the study, they used different datasets composed of MRI, T1, and T2 contrast-enhanced images, as well as FLAIR scans for each patient, depicting a knee. Contrary to the baseline method, results suggest that the proposed pipeline can provide much better cross-modality results in terms of accuracy.

In [150] the authors confirm the greater efficiency of DL models over (compared to) handcraft feature-based approaches. Hence, DL models outperform conventional methods. Specifically, the authors proposed a comparison between handcrafted features approaches and DL methods with a multi-label strategy.

Cheng et al. [158] developed a cascaded CNN to extract discriminative features from medical images of the ADNI dataset to distinguish MRI from PET images. The overall results indicate that the application of cascaded CNN further improves classification accuracy.

The authors of [159] proposed a deep architecture for exploiting multi-modality data. More in detail, the authors performed feature extraction from a set of ROIs related to MRI and PET scans from the ADNI dataset. The proposed multi-layered architecture composed of several autoencoders classified the images, achieving good results.

Arias et al. [160] applied a discrete Bayesian network to discriminate medical images from *ImageCLEFmed 2013* collection, which is a dataset containing more than

2,000 annotated images for both training and test set. The total number of classes is 31. The authors performed a low-level visual features extraction (i.e., Bags of Color, Fuzzy Color, and Texture Histogram, etc.). Since the dataset contains many classes, the authors proposed a hierarchical approach to deal with the multi-class problem. They partitioned the original problem into recursive sub-problems by exploiting the intrinsic relationships between class values. Finally, they used the proposed *Averaged One Dependence Estimators* (AODE) method, based on Naive Bayes classifier, to distinguish images with different modalities. Results confirmed that the proposed method outperforms other classification models when a hierarchical approach is used.

3.4.3 Dataset

Data used in this study were downloaded from ADNI, a multi-site study partnership launched in 2004. The acronym ADNI stands for Alzheimer's Disease Neuroimaging Initiative. This study is the result of joint effort in 2003 between the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the Food and Drug Administration (FDA). The primary goal was to verify whether clinical, imaging, genetic and biochemical biomarkers are effective in clinical trials. More details can be found on the ADNI website[3]. In this dissertation, we used NIFTI images from the ADNI study. In particular, data of MCI, Patient, AD research groups were downloaded from ADNI as both training and test sets. For the purpose of our study, we selected medical images that have following modalities:

- Diffusion Tensor Imaging (DTI)
- Functional Magnetic Resonance Imaging (fMRI)
- MRI (T1 and T2)
- PET

Each modality includes more than 6000 exams (although there are multiple exams from the same patient). The class (modality) with the highest number of images is the one performed with T1-weighted parameters (more than 21000 cases). The examinations are related to patients aged 55 up to and including 90 year olds. The enrolled participants had a clinical/cognitive assessments for approximately 2/3 years at regular intervals (6 or 12 months). The total number of AD participants was 200. All subject have been studied at specific intervals (0, 6, 12 and 24 months). Similarly, MCI patients have been analyzed at 0, 6, 12, 18, 24 and 36 months. Clinicians performed both MRI and PET scans. If quality issues have been raised, a process of re-scanning have been performed. More details are reported on the ADNI website [4].

3.4.4 The proposed approach

Modality classification was performed by using two types of DL models, reported as follows.

- Pre-trained models

[3] `http://adni.loni.usc.edu/` (accessed on 5 March 2021)

[4] `http://adni.loni.usc.edu/wp-content/themes/freshnews-dev-v2/documents/mri/Non_MRI%20Specific_ADNI_GeneralProceduresManual.pdf`

- NFNets model

Pre-trained model. To perform image modality classification, we firstly used two deep models: VGGNet-16 [161] and ResNet101 model [162]. Specifically, we applied a transfer learning technique that has been realized by using ImageNet weights. The performance of both models was evaluated on our dataset composed of multi-modal ADNI images subdivided into five different classes: DTI, fMRI, MRI _T1, MRI _T2, and PET. We divided the dataset into a training set of 8,500 images (1,700 images for each class) and a test set including a total of 1,500 images (300 images for each class).

VGGNet-16. The VGGNet-16 architecture includes 16 layers. In order to perform transfer learning, we firstly implemented a transition from high-dim features (4096) to low-dim features (2048). Then, we added another transition from 2048 to 512 features, followed by a Dropout layer. All the hidden layers use ReLU as its activation function. Then, the last feature vector is passed to the softmax layer to perform the classification of 5 classes.

ResNet101. ResNet101 consists of 101 layers. In our model, we used the convolution layers from Layer 0 to Layer 4 of ResNet101. Similarly to VGGNet -16, we defined a transition about high-dim features (2048) to low-dim classes (5). In fact, the feature maps of Layer 4 are fed into a Global Average Pooling (GAP) layer to generate a feature vector $f_1 \in R^{2048x1}$, followed by a Dropout layer and a ReLU layer. Finally, the feature vector is then transformed to another feature vector $f_2 \in R^{512x1}$ via a fully connected (FC) layer.

Implementation details. All experiments were conducted using PyTorch [163] over a cluster of GPU NVIDIA T4. The input images were resized to 256×256 pixels and randomly cropped by 224×224 pixels. A random horizontal flipping is applied for data augmentation for training data. We use Stochastic Gradient Descent (SGD) [164] optimizer and batch normalization as the regularizer. We train the model for 100 epochs. The batch size was set to 16. Moreover, we used a weight decay of 0.0005 and a momentum of 0.9.

NFNets: Model without Batch Normalization.

Since the introduction of DL networks, batch normalization has been a widely used technique to normalize the data of each batch. In fact, this strategy has always provided advantages not only to smooth the loss but also to solve the problem of *internal covariant shift*. Batch normalization permits to prevent the change of distribution related to the inputs between the hidden layers. The main issue of the neural networks concerns the effect of the initial weights selection: DL networks are somewhat sensitive to them. Motivated by this issue, batch normalization has been proposed as a valuable method to shift the inputs according to a calculated mean and variance, which reduce the loss function and stabilize the training process. However, batch normalization not only has pros but also cons, which have prompted researchers to design alternative solutions for training neural networks. Over the past year, the *Google*'s DeepMind[5] research group has developed a new family of neural networks to perform image classification. This new type of neural network, called Normalizer-Free Networks (or NFNets) [165], has outperformed previous state-of-the-art methods, achieving a level of accuracy on ImageNet equal to that of EfficientNet-B7 [166]. The most interesting new feature is the training phase, which is 8.7x faster than previous methods [165]. As previously mentioned, it is a common practice to scale and center the data on the same mean and variance by using batch normalization. Despite being effective, batch normalization poses

[5] `https://deepmind.com/`

some issues that have affected the performance of DL models. Firstly, batch normalization requires an expansive computational power due to the computation of the mean and variance for each batch, which must be stored for back-propagation. Secondly, there exists a discrepancy between the data processing in the training and testing phase. During training, data are generally processed considering a given number of batches. However, if a single instance of entities is provided during the testing phase, this process may lead to low performance. Finally, batch normalization destroys the independence between data in the same batch, implying two consequences:

- the size of each batch is a crucial factor. A small batch size leads to an approximate average rather noisy. On the contrary, a better generalization is provided by choosing a larger batch, guaranteeing a more stable training.
- The distributed training is inefficient. Distributed training reproduces a training phase on parallel streams, which are processed independently. In this regard, a discrepancy is observed when there is no communication between batch normalization layers, i.e., mean and variance parameters are calculated separately. This process leads to a major drawback as the processed data do not refer to the entire batch but only to a small portion. This issue does not allow the calculation of the loss function correctly.

Motivated by these issues, the *DeepMind* group has developed a novel neural network without implementing batch normalization in order to enhance the performance and speed of training. Specifically, the main novelty consists of a new function called *Adaptive Gradient Clipping*, which allows to "cut" the gradient if its values exceed a given threshold.

Formally, we define the Adaptive Gradient Clipping formula as follows.

$$G_i^l \rightarrow \begin{cases} \lambda \frac{max(\|W_i^l\|, \epsilon)}{\|G_i^l\|} G_i^l, & \text{if} \frac{\|G_i^l\|}{\|W_i^l\|} > \lambda \\ G_i^l, & \text{otherwise} \end{cases} \tag{3.4}$$

where G_i^l denotes the i^{th} row of the gradient matrix G^l, W_i^l indicates the i^{th} row of the weights matrix W^l, and ϵ refers to a small constant of 10^{-3}.

Generally, the standard *Gradient Clipping* function allowed the gradient to assume large values in both positive and negative directions. Basically, we aim to avoid the large "jumps" of the gradient while searching for the local minimum. The standard *Gradient Clipping* formula is reported as follows:

$$G_i^l \rightarrow \begin{cases} \frac{G}{\|G\|} \lambda, & \text{if} \|G\| > \lambda \\ G & \text{otherwise} \end{cases} \tag{3.5}$$

In order to reduce the dependency on the clipping threshold value (λ), the authors implemented the *Adaptive Gradient Clipping* function. In this regard, the "cut-off" of the gradient depends on the ratio of the norm of the gradients to the norm of the weights of a given layer ($\frac{\|G_i^l\|}{\|W_i^l\|}$). This ratio suggests how much the gradient affects its initial weights.

More details regarding the NFNets architecture are reported on [165].

Implementation details. Pytorch [163] DL library was used to perform the experiments detailed below. We have run our experiments as well as training and testing of the proposed DL architectures in Google Colaboratory (Colab) environment running in a server having a Tesla K80 Graphics Processing Unit (GPU). We

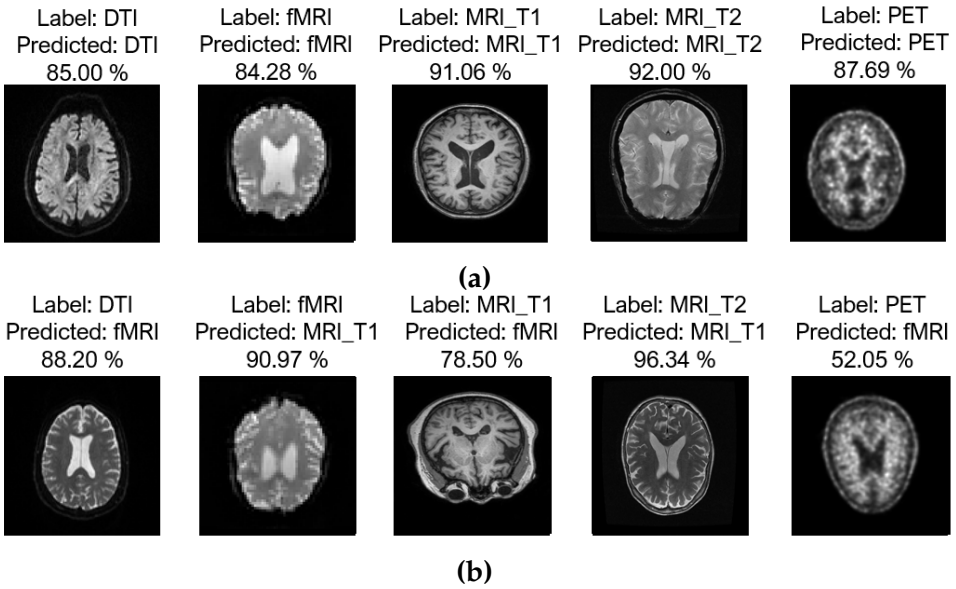


Figure 3.3: Example of image classification performed by NFNets. (a) good classification and (b) bad classification. Each raw includes 5 examples of each involved medical images correctly or not classified together with its confidence score.

resized input images by 256×256 pixels. A 224×224 center crop is sampled from an augment image, applying geometric transformations. The network is trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9. We set initial learning rate to 0.01, decaying learning rate by a factor of 0.00002 every 7 epochs. We set the number of epochs to 100. All the experiments use a batch size of 64.

3.4.5 Results and Discussion

NFNets achieved a classification accuracy of 89.6%. Results suggest that the network is able to classify MRI modality images with T1 and T2 weights correctly. Instead, it tends to misclassify more DTI, fMRI, and MRI with T2 weights classes. In addition, the low confidence score related to fMRI images depends on the fact that they belong to the least numerous class, thus processing fewer examples in the training set. In Fig. 3.4.5, examples of modality classification performed by NFNets are depicted. Regarding pre-trained networks, the VGGNet-16 network reports an accuracy equal to 94.7%. It correctly classifies most classes with good confidence, although it has more difficulty identifying fMRI and PET class images. Furthermore, it tends to misclassify DTI with fMRI as they present similar characteristics. VGGNet-16 also erroneously classify fMRI as PET images with a high confidence score. In Fig. 3.4.5, we reported some examples of classification performed by VGGNet-16. On the contrary, the ResNet101 network classifies the PET class with higher accuracy than VGGNet-16, whereas it identifies the fMRI images as MRI _T2 images. Although these results represent a first baseline, note that they are just preliminary. In Fig. 3.4.5, we reported examples of both correct and incorrect classification performed by the pre-trained ResNet101. The study conducted represents a preliminary approach to apply Deep Learning techniques for medical image modality classification. This work is currently under development, intending to define a more innovative approach to enrich the state-of-art. In this study, the difficulties encountered relate mainly to the

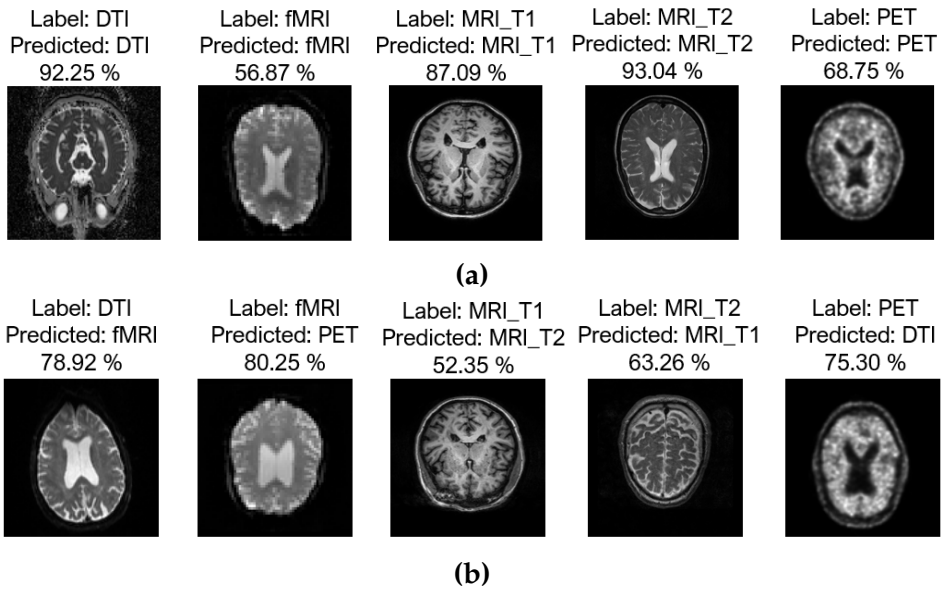


Figure 3.4: Example of image classification performed by VGGNet-16. (a) good classification and (b) bad classification. Each raw includes 5 examples of each involved medical images correctly or not classified together with its confidence score.

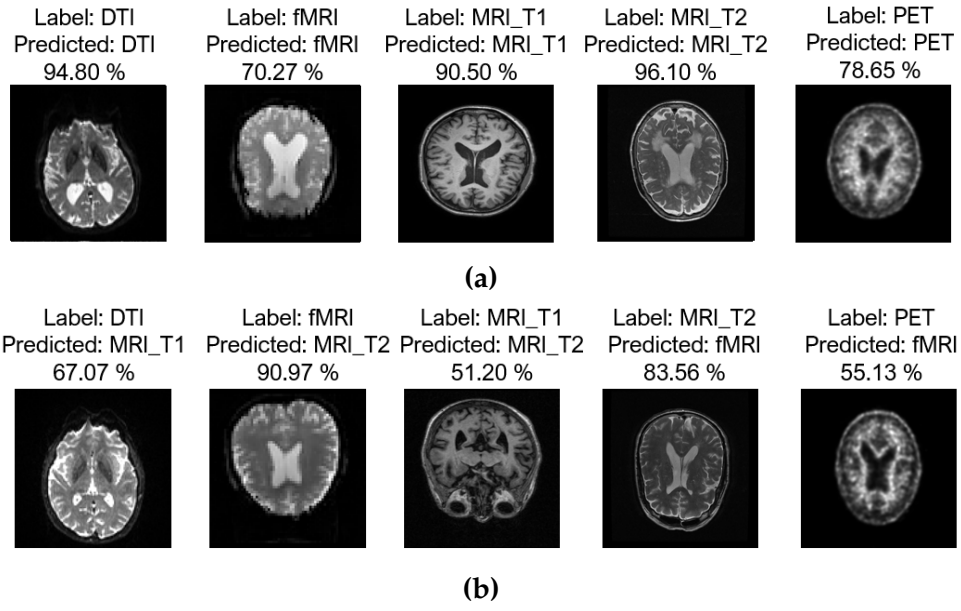


Figure 3.5: Example of image classification performed by ResNet101. (a) good classification and (b) bad classification. Each raw includes 5 examples of each involved medical images correctly or not classified together with its confidence score.

quality of the dataset that, initially, presents images with low quality (especially with regard to PET). In addition, the present dataset has a limited number of data due to an initial skimming carried out in the preliminary phase of this study that led to the removal of several images from the dataset, as they presented a poor quality and did not have a suitable format for our experiments. As anticipated, these issues are pretty common in the medical imaging domain. Nevertheless, the results obtained encouraged us to define a more in-depth study.

3.4.6 Future Works

As anticipated, the results shown represent a first insight on the modality classification task. Considering this, we aim to deepen this investigation by applying DL networks to perform an even more accurate and efficient image modality classification. In this regard, our future purposes include implementing more advanced pre-trained neural networks, such as EfficientNet, to evaluate their performance and provide accurate benchmarking. In addition, we aim to provide a more significant contribution by defining a new loss function that can effectively improve image classification. Finally, we aim to apply *Transformer networks* [167] which recently have been emerged as the new efficient paradigm. In general, Transformer networks are used to solve several problems, ranging from speech recognition problems to text-to-speech transformation, i.e., problems of sequence transduction. Transformer Networks are similar to RNN and LSTM networks. Contrary to these recurrent networks, Transformer networks adopt an attention strategy to focus on each element of a sequence to predict the output, leading to impressive results [167]. In this regard, our future work will focus on applying these networks to our classification problem to boost image classification accuracy.

3.5 COVID-19 Disease Segmentation

3.5.1 Overview

In late 2019, Chinese health authorities notified an outbreak of fatal etiology pneumonia in Wuhan, China, killing hundreds of people. The virus has been named Corona Virus Disease (or CoVID-19) by the Worldwide Health Organization (WHO). However, the official name is *Severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2). Rapidly, the virus spread across Europe and eventually spread around the world. On March 11, 2020, the WHO declared that the COVID-19 outbreak could be considered a pandemic. Specifically, COronaVIrus Disease 2019 (COVID-19) is an acute respiratory infection of SARS-CoV-2. The main clinical symptoms of this disease include fever and fatigue, while respiratory symptoms include a mainly dry cough and shortness of breath [6]. In severe cases, this syndrome causes respiratory distress, which may even lead to death. By analyzing CT scans related to thoracic imaging, clinicians observed that the presence of the virus is indicated by bilateral pneumonia of an interstitial type. The symptoms of COVID-19 vary according to its severity: ranging from no symptoms (being asymptomatic) to presenting fever, cough, sore throat, weakness, fatigue, and muscle pain. In many cases, clinicians observed that it could lead to pneumonia, acute respiratory distress syndrome, and other complications. In this challenging scenario, researchers have quickly demonstrated that processing images from CT scans can provide a valuable diagnosis to

[6] https://www.ecdc.europa.eu/en/covid-19/latest-evidence/clinical

assist experts in the medical field. Therefore, pulmonary CT or chest X-ray play a fundamental role in the diagnosis phase. Although recent studies have shown that CT is characterized by low sensitivity in identifying the earliest pulmonary changes in COVID-19, CT remains a valid tool in the evaluation of the presence of pneumonia and the progression of the disease. Compared to X-ray examinations, CT scans are also more helpful given the sharper level of detail. In addition, the presence of COVID-19 is characterized by peculiar elements (e.g., *ground glass*, lung consolidations, etc.). Although they can provide useful insight for a preliminary visual classification, their identification is not sufficient to confirm the presence of disease infection. In general, the predominant clinical manifestations of COVID-19 are:

- Presence of the disease in both lungs, often in the form of opaque spots
- Peripheral and bilateral *ground glass* opacity (GGO), with or without consolidation or *crazy-paving*.

Other less common features include the presence of sporadic GGOs with a non-rounded and non-peripheral distribution.

Since the spread of the emergency situation, a decisive contribution has come from AI: in fact, since the COVID-19 epidemic is spreading rapidly and on a global scale, it is essential to have full knowledge in order to deal with it. As already discussed, AI can be very useful in the medical field, as it can assist professionals in classifying and retrieving patient conditions, understanding why and how patients develop diseases, considering what treatment will be most suitable for them. In this regard, the most used techniques are based on ML and CV algorithms. In this section, we investigate the problem of developing effective solutions for the segmentation of lung infections. More specifically, the purpose is to lay the foundations for the definition of an advanced system that is able, by observing the CT exams of a patient affected by COVID-19, to establish with a high degree of confidence the presence or absence of the pulmonary infection. By analyzing a set of clinical histories related to patients affected by the virus, we aim to define the outcome of disease and progression. This work is intended as a result of the collaboration with the team of the Department of Medical and Surgical Sciences and Advanced Technologies, led by Dr. Stefano Palmucci, a radiologist at "Azienda Ospedaliero-Universitaria Policlinico-Vittorio Emanuele" of Catania.

The preliminary steps of this research can be summarized as follows:

- Subdividing the lung region into its five lobes;
- Identifying COVID-19 infection using advanced several neural networks;
- Analyzing data in order to detect recurrent patterns regarding the percentage of disease concentration in the lung lobes;
- Evaluating results, using metrics that compare them and the labeling provided by clinicians.

3.5.2 State of the Art

Since the spread of the COVID-19 disease, the scientific community has been spent so much effort to develop effective solutions for helping clinicians at determining the presence of this infection through medical image analysis. In this scenario, most of the solutions propose the use of advanced CV and DL techniques. In [168], several

DL methods to distinguish coronavirus disease from pneumonia and normal classes were applied. The proposed method used different DL models (such as VGGNet, DenseNet, etc.) for the classification of Chest X-Ray. Due to the non-availability of large medical imaging dataset, the authors applied the transfer learning technique. Results reported that the pre-trained Inception _Resnet_V2 achieved the highest values in terms of Specificity, Sensitivity, Precision and Accuracy, whereas DenseNet201 obtained good metrics values. On the contrary, VGG19 reports the lowest values. Chen et al. [169] proposed a system to detect Covid-19 disease on CT-scans. After collecting more than 40,000 CT images from 106 patients, the authors filtered the images with good lung fields to be used for the training and test step. The enrolled radiologists annotated infection lesions of COVID19 pneumonia. The authors used U-Net++, a novel architecture, to perform the image segmentation. The pre-trained ResNet-50 model was used as backbone. Specifically, all the pre-training parameters of ResNet-50 are loaded to UNet++. The overall results confirmed the effectiveness of the proposed solution at detecting COVID-19 disease. Similar to previous work, the authors of [170] proposed a classification system which combines the predictions of several DL approaches (such as VGGNet-16, ResNet101, etc.). The decision fusion system is based on the majority voting approach. The results show that the proposed method achieve 86% in terms of Accuracy, Sensitivity, Specificity, F1-score, Precision and Recall, confirming its effectiveness in discriminating COVID-19 on CT-scans. In [171], the authors implemented a novel system for the classification of COVID-19 using CT-scans. The overall pipeline consists of five steps. The first one is devoted to the data acquisition, the second consists of the application of CNN model, the feature extraction is used in the third step, whereas the choice of more robust features is performed in the fourth one. Finally, the fifth step passes the extracted features to One-Class Kernal ELM (ELM stands for extreme learning machine) to perform the classification. Chao et al. [172] proposed a study to combine radiomics of lung opacities and non-imaging features from demographic information to predict whether a patient is eligible for intensive care unit (ICU) admission. The authors selected features including hierarchical lobe-wise quantification (HLQ), whole lung radiomics (WLR) features as well as demographic, vital signs, and blood examination features (DVB), combining them through the use of a fusion strategy. Moreover, they implemented a lung lobe segmentation approach in order to segment both lungs, five lung lobes and pulmonary opacities from non-contrast chest CT examinations. The model used to perform the segmentation were proposed by [173] and [174]. Finally, a Random Forest (RF) classifier is applied to predict ICU admission. Despite being a promising approach, its overall performance is not exceptional due to the limited size of the data.

3.5.3 Lung composition

The lungs are the two organs responsible for the supply of oxygen to the body and the elimination of carbon dioxide from the blood, i.e., the gaseous exchange between air and blood (a process known as hematosis). The thoracic cavity includes the lungs, which are enveloped by a serous membrane known as the pleura. The right lung is composed of three lobes (upper, middle, and lower) separated by an oblique and a horizontal fissure, while the left lung consists of two lobes (upper and lower) separated by an oblique fissure. The lobes are further subdivided into bronchopulmonary segments, each of which is served by a segmental bronchus; the segmental bronchi, in turn, are subdivided into smaller and smaller structures until arriving at

the pulmonary alveoli, the structures responsible for gaseous exchanges between air and blood.

3.5.4 The original dataset

The data collection consists of DICOM-format images produced by CT scans from 45 patients affected by COVID-19. The enrolled patients were being treated at "Azienda Ospedaliero-Universitaria Policlinico-Vittorio Emanuele" of Catania. For research purposes, all CT examinations were appropriately anonymized by clinicians. For each patient, an average of 250 slices is reported. Each slice has a resolution of 512×512 pixels, depicting (in section) the thoracic area. Clinicians have indicated different scores according to the amount of COVID-19 observed in an individual lobe, defining the following five scores:

- 1 - Range from 0% to 5%;
- 2 - Range from 6% to 25%;
- 3 - Range from 25% to 50%;
- 4 - Range from 51% to 75%;
- 5 - Range from 76% to 100%;

where label 1 indicates a low presence of COVID-19, whereas label 5 denotes a rather severe COVID-19 pneumonia. Clinicians provided us a document containing the following information related to each patients:

- Col 1 - ID Patient;
- Col 2 - The first letter of the name and the last name for each patient;
- Col 3 - Date of birth;
- Col 4 - RUL (i.e., Right Upper Lobe), indicating the amount of concentration of COVID-19 infection for the lobe;
- Col 5 - LUL (i.e., Left Upper Lobe);
- Col 6 - ML (i.e., Medium Lobe);
- Col 7 - RLL (i.e., Right Lower Lobe);
- Col 8 - LLL (i.e., Left Lower Lobe);
- Col 9 - Total Score, indicating the sum of each rows where labels for each lobe are reported;
- Col 10 - Notes about qualitative patterns (such as, ground glass or consolidation)

We encountered some challenges in processing the images during the research investigation due to the non-adequate data labeling and the unbalanced dataset. Firstly, we observed that the patient identifiers used in the shared document did not match or could not be reported in the dataset. Annotations are indicated by the initials of the patients' names (e.g., CP). However, since there are several homonyms

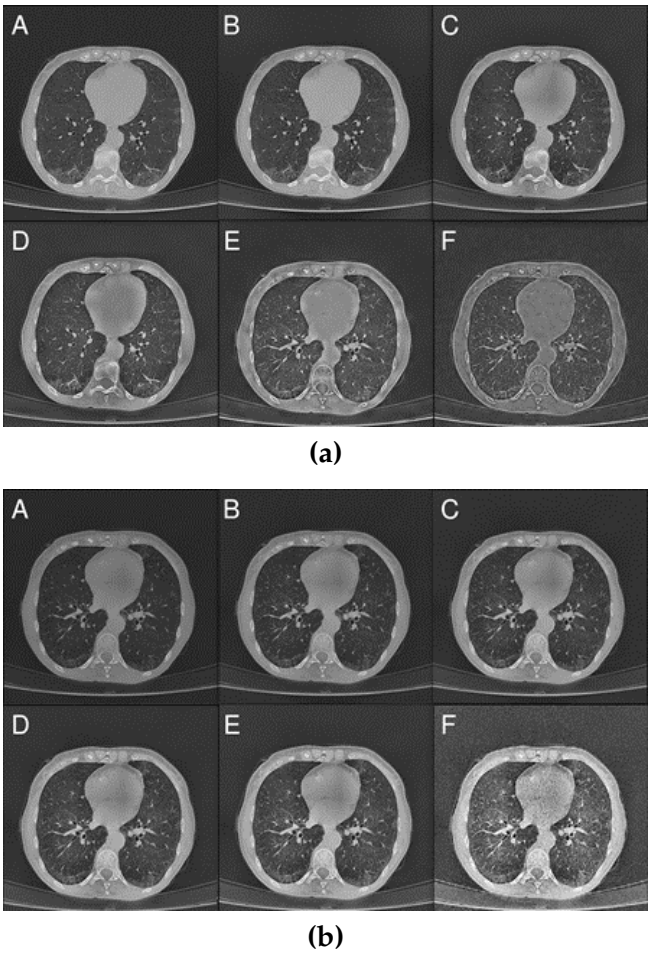


Figure 3.6: Examples of images after applying CLAHE function. (a) *tilGridSize*, (b) *clipLimit*.

and the date of birth is not provided, it has been challenging to find the corresponding patient in the dataset. Motivated by these findings, we performed a preliminary skimming of the cases to be considered for our experiments. Secondly, there is a substantial imbalance in the data as the most numerous class is related to label 2 (more than 18 cases). On the contrary, label 5 includes a few cases (5). The imbalanced dataset could affect the robustness in the final instance. Finally, clinicians provided only labeling regarding the concentration of infection present in each lobe of the lungs. However, they did not perform labeling to detect COVID-19 disease. The absence of labeled data poses a major challenge in defining a strategy for the segmentation of this disease. In the following sections, we explain how the lack of labeled data has hindered our research activity.

3.5.5 The pre-processed dataset

In order to improve the performance of the applied algorithms, we performed a preprocessing operation using the CLAHE function. The latter is devoted to equalizing the histogram of an image in order to balance its contrast. The CLAHE function allows distributing the contrast over the whole histogram. Contrary to other equalization techniques, it does not work simultaneously on the whole image but acts locally, focusing on image fragments known as *tiles*. Each tile is a non-overlapping

Parameters	**Values**					
clipLimit	1.0	2.0	3.0	4.0	5.0	10.0
tileGridSize	$(2,2)$	$(4,4)$	$(8,8)$	$(16,16)$	$(32,32)$	$(64,64)$

Table 3.2: Selected values for *clipLimit* and *tileGridSize* parameters.

piece of a given image. Once the maximum contrast value (Contrast Limited) has been defined, the pixels that exceed this value are "cut" in order to make the histogram more homogeneous. To determine the optimal values of *clipLimit* and *tileGridSize*, we chose different values for each parameter, combining the values to provide possible combinations. In Table 3.2, we reported the values for each parameter. In Fig. 3.6, examples of images considering the selected values are depicted. Note that, as the parameters increase, the noise is more evident within the images. It can be observed that outlier cases lead to processing issues, whereas average ones provide a good compromise between image quality and resolution details which allows detecting the pulmonary infection efficiently.

3.5.6 COVID-19 Segmentation

The approach used is based on networks called Inf-Net, an architecture recently introduced by Fang et al. [175] for the segmentation of pulmonary infections, including COVID19. The used DL model was firstly evaluated on the original dataset. Then, a subset of the dataset processed by Contrast Limited Adaptive Equalization (CLAHE) function was defined, which will be further explored in the following sections.

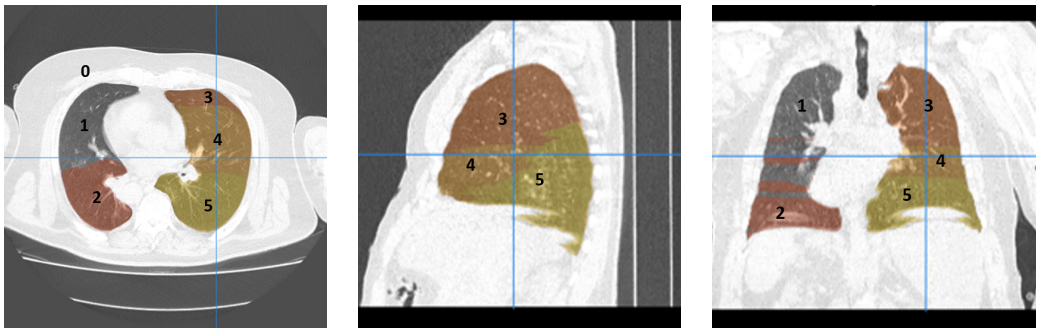
The automatic detection of lung infections from CT images has offered great potential for defining an effective healthcare strategy. However, it has been found that the infected regions, observed by CT examination analysis, present a high variation that makes disease identification harder. In order to overcome this issue, several techniques based on DL algorithms have been proposed in order to perform the segmentation of the infected regions. As previously mentioned, we adopted a DL model, Inf-Net, which was pre-trained on COVID-19 CT Segmentation Dataset[7] to cope with the absence of labeled data in our dataset.

3.5.7 Lobes Identification

Before proceeding with the detection of the COVID-19 disease, it is required to identify the five lobes from CT analysis. For this purpose, we adopted a U-net-based neural network, named U-net(LTRCLobes_R231) [173]. The latter combines, in turn, two types of neural networks which operate on different fronts: U-Net(R231) and U-Net(LTRCLobes). The former deals with the recognition of lung contours from CT analysis. On the other hand, the second model subdivides the lungs into five regions (lobes) by means of scissure recognition. Consequently, the U-Net(LTRCLobes_R231) merges the two mentioned approaches to better define lung contours and lobes. The identification of the lobes is performed by assigning each pixel in the image a value between 0 and 5, which identify the following areas respectively:

- 0 - region outside the lungs
- 1- left upper lobe

[7] https://medicalsegmentation.com/covid19/



(a) (b) (c)

Figure 3.7: Examples of lung lobes segmentation from different points of view.

- 2- left lower lobe
- 3 - right upper lobe
- 4 - middle lobe
- 5 - right lower lobe

In Fig. 3.7, an example is shown.

3.5.8 Inf-Net architecture: a brief overview

The Inf-Net neural network was proposed by Fang et al. [175] to automatically identify the COVID-19 infection from CT scans. In Fig. 3.5.8, we report outputs generated by the Inf-Net.

The network takes input images deriving from CT scans and processes them through convolutional layers to extract predefined features. In detail, the first two convolutional layers extract low-level features analyzed by an attention module, called *Edge Attention Module (EAM)*, which ensures that the extracted features are within the boundaries of the region of interest. Then, the collected features are processed by three other convolutional layers to extract the high-level ones. The latter is first used by a Parallel Partial Decoder (PPD) to aggregate these features and generate a global map for coarse localization of lung infections. Then, the combination of the extracted features is taken as input by cascaded Reverse Attention (RA) modules that allow the final prediction of lung infection within the thoracic region.
The Inf-Net architecture consists of several modules, reported as follows:

Edge Attention Module. This module allows us to extract information about the edges of a specific region of the image, taking low-level features as input to create a so-called edge map. The edge map consists of an image that indicates the different edges in the source image, allowing the identification of the region of interest from which to extract information. In this case, the analyzed region is the thoracic one, in which we determined the presence or absence of COVID-19 infection. Once the edge map is obtained, we compare the edge map produced by this module and the one relative to the ground truth image using the loss function, which indicates the error rate.

Parallel Partial Decoder. In medical applications based on Machine Learning techniques, the processing of high and low-level features is often required. However, the high computational cost required by the latter due to the larger spatial

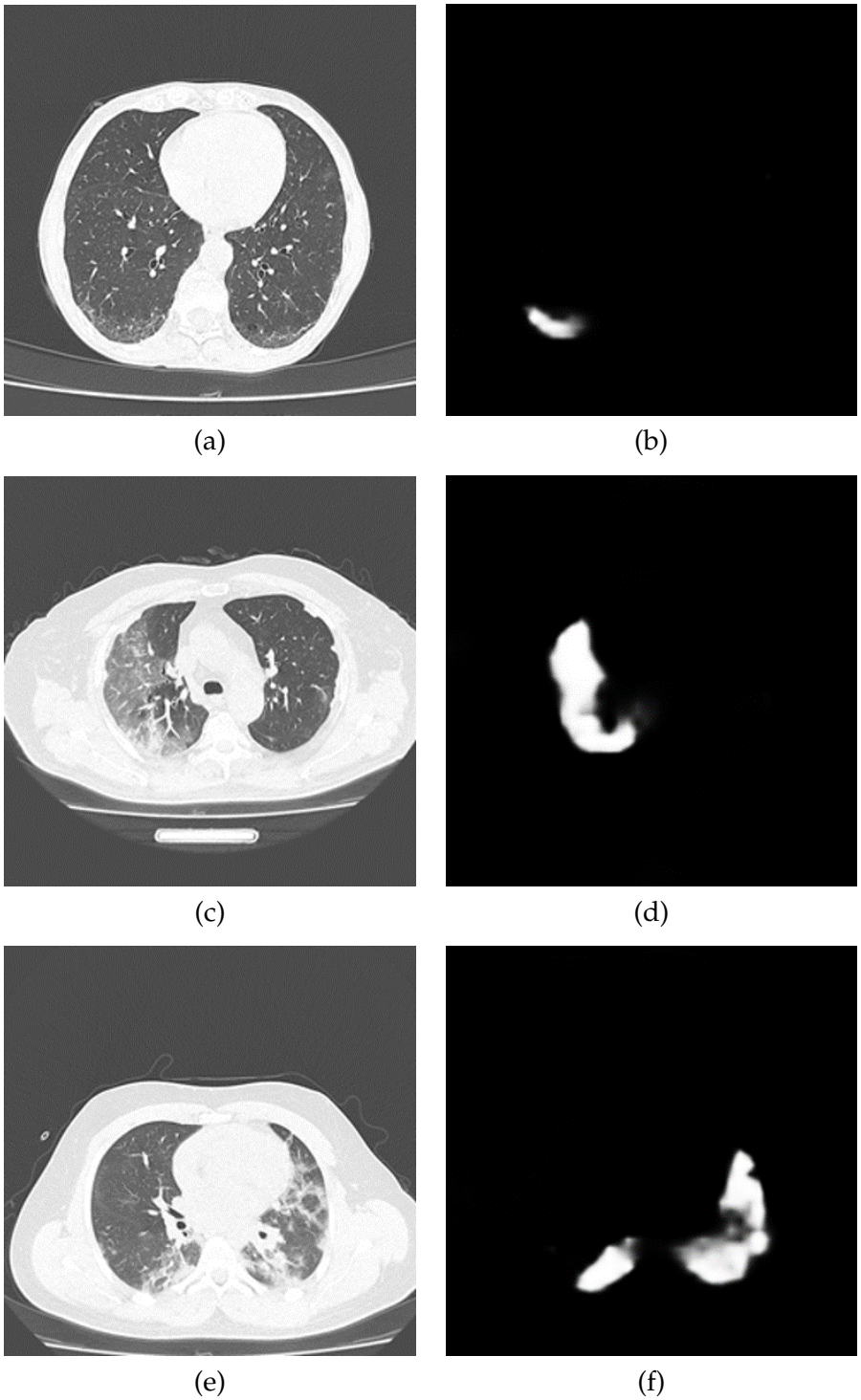


Figure 3.8: Output results of COVID-19 disease segmentation. (a) Original slice of a CT scan, (b) Mask produced by Inf-Net.

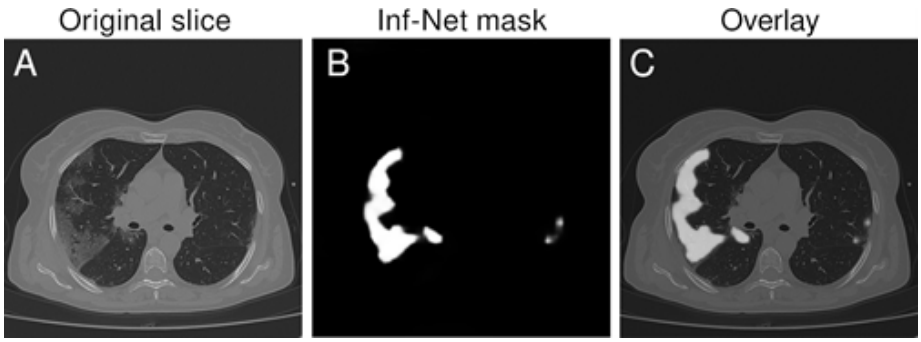


Figure 3.9: A: Original slice of a CT scan; B: Mask produced by Inf-Net for the slice in subfigure A; C: Overlay of A and B.

resolutions and the small contribution made to the performance raises some doubts about its use. For this reason, the method proposed by [175] uses a Parallel Partial Decoder for the processing of high-level features. In particular, for each input image, two sets of low-level features and three sets of high-level features are extracted using five levels of convolution. Finally, the partial decoder allows to aggregate the high-level features and create a global map representing a sort of guide for the subsequent Reverse Attention Modules (RA).

Reverse Attention Module. In the clinical setting, the procedure for segmentation of infected regions consists of two main steps: localization of the region of interest and labeling of the region of interest by observation of lung tissue structure. The Inf-Net makes use of two different components that operate as a localizer and a labeler. As anticipated, the PPD component allows to locate the region of interest by creating an appropriate edge map, and a progressive framework is used to label the data. Specifically, the authors implement reverse attention modules that adaptively learn high-level features.

More details about the Inf-Net model are provided on [175].

3.5.9 The proposed application

Given the absence of labeled data, the pre-trained Inf-Net on an additional clinical dataset was used for the purposes of this study. For each patient of the dataset provided by Dr. Palmucci's team, each single slice has been extracted from the data in DICOM format, obtaining as output images in PNG format. Subsequently, the experiments were conducted through the use of the pre-trained Inf-Net, which produces in output the masks representing the segmented regions, in grayscale, in which the presence of COVID-19 was detected. An example of a mask produced by the used network is reported in 3.9.

3.5.10 The post-processed dataset

After applying the Inf-Net model, the resulting masks presented artifacts that could have affected the performance of the detection disease. For this reason, we performed a post-processing operation, applying thresholding values in order to produce denoised masks. Then, the morphological opening operator is applied in order to remove smaller artifacts caused by filters or compressions.

3.5.11 Lobe Filling: a strategy to evaluate the amount of COVID19

Algorithm 1 Lobe Filling algorithm.

Initialize a couple of variables $count1_k, count2_k, \forall k = 0, \ldots, 5$.
for slice k in Set_{Inf} **do**
  for (x,y) in slice $S_{Inf_k}(x, y)$ **do**
    if $S_{Inf_k}(x, y) > \phi$ **then**
      if $1 \leq S_{Lobe_k}(x, y) \leq 5$ **then**
        $count1_k \leftarrow count1_k + 1$
      else
        $count2_k \leftarrow count2_k + 1$
      end if
    end if
  end for
end for
Computing the percentage.
Normalize values according to the rules defined by clinicians

In Lobe Filling algorithm 3.5.11, we report the pseudo-code of the proposed method for estimating the amount of infection on each lobe.

Once the variables $count1_k, count2_k$ are defined, $\forall i = 0, \ldots, 5$, where the values of i denote the lobes of the lungs, a comparison is defined between a Set_{inf} image (relating to infection segmentation) and a Set_{lob} image (relating to lobe segmentation). Both slices referred to the same patient. The comparison is performed by defining an iteration for each slice of the Set_{inf}, checking the following conditions:

- Whether the pixels under consideration are part of the "infectious disease zone". We check whether the pixel value is equal to 255 or less (non-black pixels indicates the presence of infection disease). - whether the spatial coordinates (x, y) of the pixel that falls within the "infectious disease zone" is part of one of the 5 lobes. Pixels with value 0 are discarded, since they identify the zone outside the lung area.

Finally, the values of the $count1_k$ and $count2_k$ counters are incremented for each identified lobe to determine, in the next step, the percentage of infection present in each lobe.

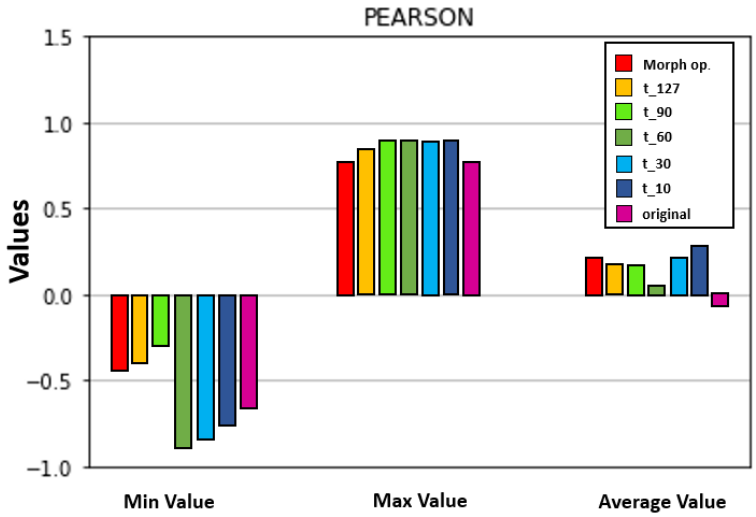
Once this percentage is calculated, it is normalized according to the labels defined by the radiologists.

3.5.12 Results and Discussion

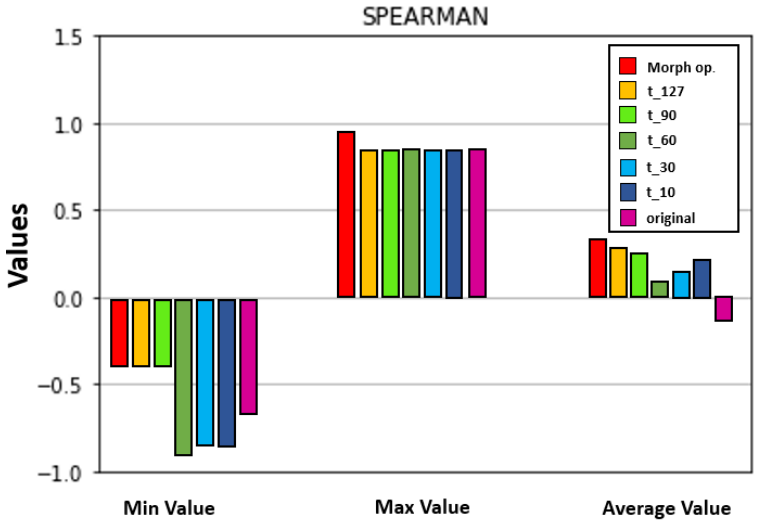
To evaluate the performance of the algorithm at segmenting COVID-19 infection, we compared the values obtained from our analysis and those defined by radiologists. To determine the correlation between the two sequences of values, we used the following coefficients:

- Pearson's Correlation Coefficient (PCC)
- Spearman's Correlation Coefficient (SCC)
- Root Mean Square Error (RMSE)

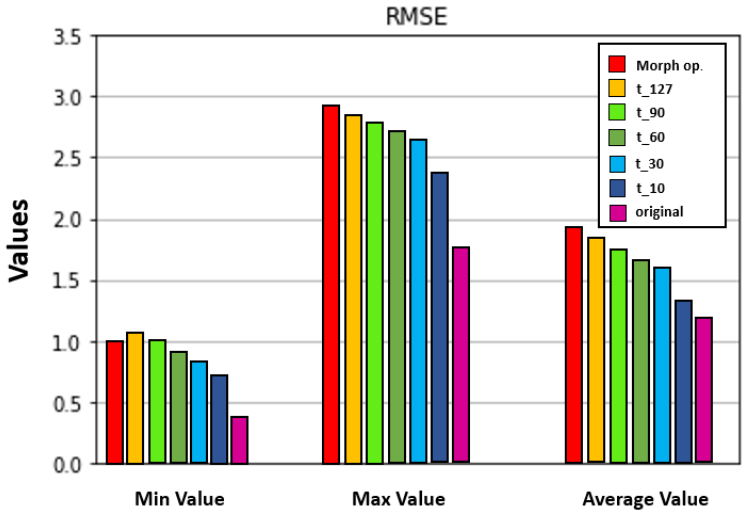
Results on the original dataset + post-processing. In Figs. 3.5.12, 3.5.12 we report the dataset evaluation based on Pearson, Spearman and RMSE coefficient. All data is grouped by thresholding level. For each group are reported the average, the worst, and the better value. With regard to Pearson, we observed that the values for



(a)

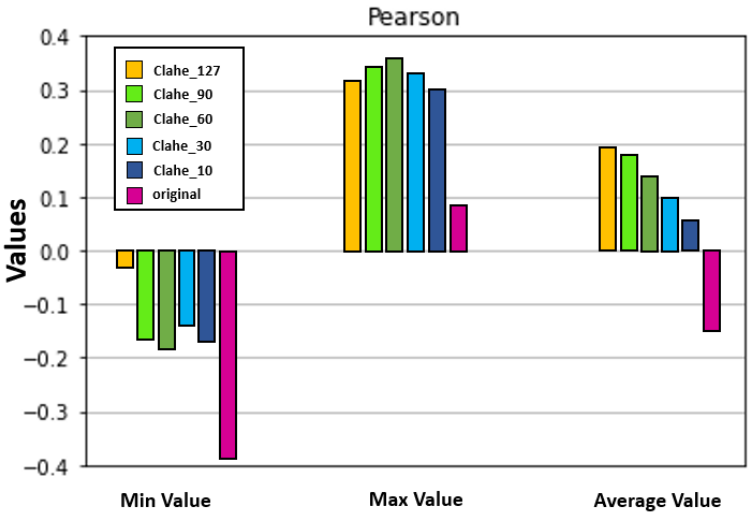


(b)

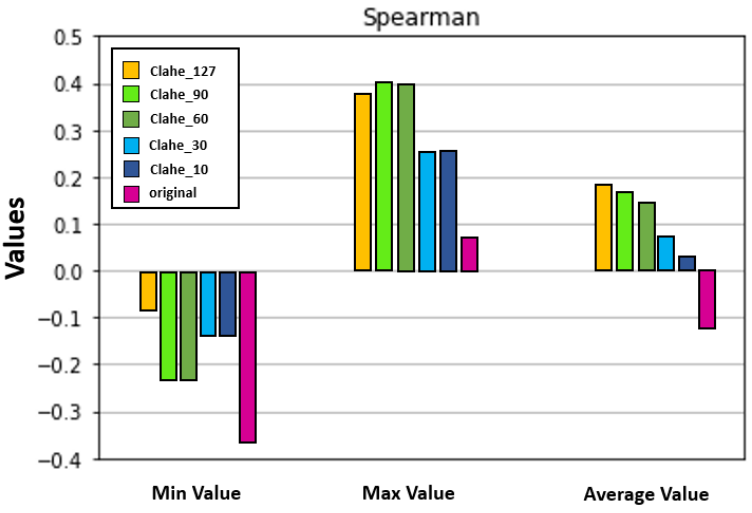


(c)

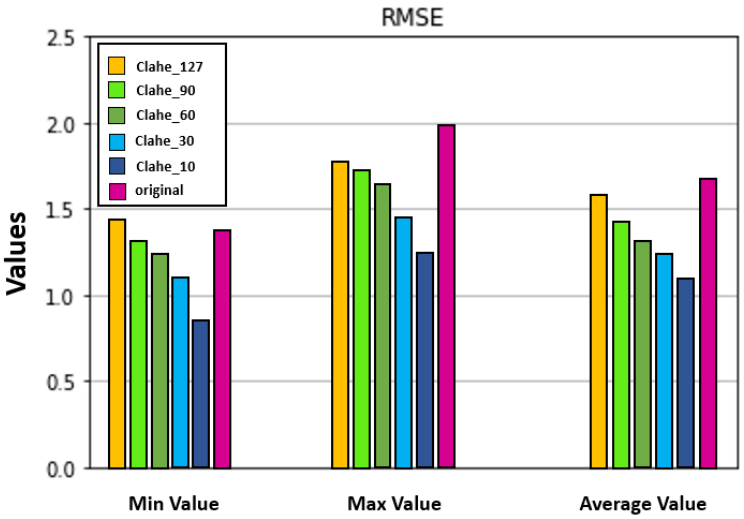
Figure 3.10: The calculated correlation metrics. (a) Pearson, (b) Spearman, (c) RMSE.



(a)



(b)



(c)

Figure 3.11: The calculated correlation metrics. (a) Pearson, (b) Spearman, (c) RMSE.

each threshold are quite similar (ranging from 0.85 to 0.95). In contrast, the worst values depend on the choice of threshold and whether or not the post-processing technique was applied to process the images obtained from Inf-Net. In general, the average value shows that the value with a higher correlation is obtained using a threshold equal to 10. Regarding Spearman, it confirms Pearson's results, although it can be noted that applying post-processing achieves better results on average, as does apply a threshold of 127 and 90. On the other hand, RMSE values are better when the masks produced by Inf-Net are used. The results are also satisfactory if a threshold of 10 is chosen. We argued that this result depends on the fact that, in the post-processing phase, the filters have removed information related to the COVID-19 infection, therefore leading to worse results. With regard to the masks determined by Inf-Net, the full information was entirely preserved, which consequently generated better values.

Results on the CLAHE dataset. In this regard, the Pearson and Spearman values are consistent with those of the previous case. However, it can be observed that their worst cases have lower values (about -0.4) compared to those of previous case (about -1.0), which leads us to conclude that the pre-processing step slightly improved the correlation between the values provided by the radiologists and those defined by our study. The RMSE values also present lower values, which confirms that the average error is also lower, and therefore, the CLAHE technique has allowed improving the quality of the masks so that they are processed more efficiently by the Inf-Net network.

3.5.13 Conclusions

To confirm the reliability of the disease segmentation, we provided the output masks to clinicians in order to be evaluated. Their response confirmed the effectiveness of DL approaches at segmenting both lobes and lung infection, although they present less accurate results in some cases. As anticipated, several cases were excluded from the dataset used for conducting experiments as they reported incomplete information. Therefore, we evaluated the performance of Inf-Net on a small subset, including eight patients out of 45. As shown, the results are not satisfactory but represent a first insight regarding the analysis of this dataset. Furthermore, the study of this dataset has highlighted several problems that are very common in the medical imaging domain. Previously, we anticipated the difficulty of applying DL techniques when mislabeled and/or unbalanced data are used, confirming how difficult it is to operate with inadequate datasets.

3.6 Immunotherapy Treatment Outcome Prediction

For completeness, we report a brief description regarding two studies based on the prediction of the response to immunotherapy treatment. Specifically, we address the problem of developing non-invasive methods that can assist oncologists in discriminating patients who respond to treatment from ones who do not. Despite dealing with limited datasets, we succeed in designing effective DL pipelines which achieve promising results.

3.6.1 Approach based on 3D Non-Local Net

Overview

Immunotherapy is regarded as one of the most significant breakthroughs in cancer treatment. Unfortunately, only a small percentage of patients respond properly to the treatment. Moreover, there are no efficient bio-markers able to early discriminate the patients eligible for this treatment to date. In order to overcome these limitations, an innovative non-invasive deep pipeline is investigated for the prediction of response to immunotherapy treatment. We propose ad-hoc 3D Deep Networks integrating Self-Attention mechanisms in order to estimate the immunotherapy treatment response from CT-scan images analysis and such hemato-chemical data of the patients.

Metastatic urothelial carcinoma (mUC), also known as bladder cancer, is the most common histological type of urinary tract carcinoma. It accounts for about 3% of all cancers [176]. It is more common between the ages of 60 and 70, and it is associated with about 165,000 global deaths every year. In the advanced stage, only approximately 5% of patients present a five-year survival [176]. The current standard first-line treatment of mUC is platinum-based chemotherapy. In terms of progression-free survival (PFS), the combination of chemotherapy and immunotherapy as a first-line treatment in mUC has determined an improvement in this domain [177]. However, significant and mature overall survival (OS) data are still expected. Immunotherapy has become the standard second-line treatment of mUC based on two phases III studies with Immune Checkpoint Inhibitors (ICIs) immunotherapeutic drugs such as Atezolizumab and Pembrolizumab, with a median OS reported of 8.6 [178] and 10.3 months [179] respectively. However, no more than about 20% and 30% of patients has a disease response with ICI in post-platinum and first-line treatment, respectively, although these responses tend to be more durable than those obtained with chemotherapy [180]–[182].

Therefore, it is a priority to identify and select those patients who can benefit from immunotherapy, although at the moment, there are still no reliable and clinically available biomarkers to properly choose patients who respond or progress with ICIs [87], [183]–[185]. We propose an innovative DL pipeline to predict response to ICIs immunotherapeutic treatment for patients with advanced metastatic bladder cancer (mUC) who have progressed following first-line platinum-based chemotherapy. The architecture of the proposed deep model is based on 3D Densely connected Convolutional Neural Network (3D-DCNN) with separable convolutions and self-attention mechanisms through non-local blocks [186]. The model processes computed tomography (CT-scan) imaging data and discriminates patients with high chances of response (complete, partial response, or at least stable disease) from those that, instead, is likely to show disease progression. We leverage the success of these models and extend them with a self-attention mechanism, based on non-local blocks, for better learning long-range dependencies among the input data (segmented CT scans cancer lesions). Experimental results show that the devised self-attention-based model leads to a better characterization of bladder cancer (i.e., the associated feature maps) and of the radiological visual features for predicting treatment outcome with respect to the state-of-the-art methods. Specifically, we contribute to the research area in automated immunotherapy outcome prediction through visual investigation of CT scans, as follows:

- We present a deep model that combines 3D densely connected convolutional layers (3DCNN) empowered with self-attention mechanisms for automatically

estimating the efficacy of bladder cancer immunotherapy treatment, purely based on CT imaging analysis.

- We investigate, through interpretability methods, such as Grad-CAM [187], what are the radiological CT visual features that most likely act as biomarkers for immunotherapy treatment outcome, thus providing potentially invaluable support to medical staff in evaluating the progress of bladder cancer. To the best of our knowledge, no method has tackled the task herein proposed from both the automated treatment outcome prediction and interpretability perspectives.

Materials and Methods

The proposed framework consists of a combination of 3D Densely Connected Convolutional layers (3D-DCNN) with Non-Local blocks [186], in order to implement a self-attention strategy devoted to improving the characterization of spatio-temporal dependencies of the neoplastic lesions in CT slices. The use of 3D deep convolutional layers is motivated by the results achieved by our previous work [188] demonstrating the correlation between the biological aggressiveness of bladder tumors with the dynamic morphological evolution of the interested CT lesions. As pointed out in bladder treatment cancer guidelines [189], not all metastatic lesions play a key role in the analysis of the progression of oncological disease. Motivated by these findings, we extend 3D convolution architectures with an implicit "attention module" to force the models to focus – through visual feature learning – only on the most significant parts of the RECIST lesion and their possible correlation. More specifically, the attention mechanism is implemented by concatenating Non-Local blocks [186] at different layers for capturing long-range dependencies at different scales. We report the problem of treatment outcome estimation as a binary classification task, i.e., the proposed model provides as output two-class probabilities: the first one denoting complete/ partial regression or stable disease (C1), and the other one for disease progression (C2). More details on the proposed architecture can be found on [190].

Dense Blocks

As previously mentioned, the 3D-DCNN includes densely connected blocks (dense blocks) with 3D separable convolution layers (both depth-wise and point-wise). Separable convolutions are adopted to improve efficiency (through a significant reduction of model parameters), while not affecting the output performance. Each dense block consists of a sequence of dense layers, including a batch normalization layer, a 3D convolutional layer with a kernel size of $3 \times 3 \times 3$ (depth-wise and point-wise separable), followed by a ReLU. Each dense block is followed by a transition down layer, aiming to half feature map dimension, and composed by a convolutional layer with a kernel size of $1 \times 1 \times 1$ followed by a max-pooling layer of kernel $2 \times 2 \times 2$. The output of dense blocks is then passed to non-local blocks.

Self-Attention through non-local blocks

Non-Local blocks have been recently introduced by Wang et al. [186], as a very promising approach for capturing space-time long-range dependencies and correlation on feature maps, resulting in a sort of “self-attention” mechanism. Self-attention through Non-Local blocks aims to enforce the model to extract correlation among feature maps by weighting the averaged sum of the features at all possible positions

in the generated feature maps [186]. In our pipeline, Non-Local blocks operate on almost each convolution layer to extract features in dependencies at multiple abstract levels for holistic morphological modeling of the input RECIST lesions. More details are reported on [190].

Classification Layer: The stack of fully connected

Once features are extracted through the combination of dense and Non-Local blocks, we obtain a one-dimension visual embedding (736×1). These features are then concatenated to blood tests and other clinical data. After combining the two sets of features, we feed a stack of fully connected (FC) layers by using them. The objective of this FC stack is to find additional correlations among the aggregated deep features and clinical data in order to enhance accuracy in assessing ICI immunotherapy treatment outcomes.

Dataset: Recruitment and data pre-processing

We collected a dataset of 43 CT/MRI scans from patients as part of a clinical study performed at a local hospital facility. The recruited patients have histologically confirmed bladder cancer (mUC) progressing after platinum-based chemotherapy and treated with an anti PD-L1 ICIs agent in the second or beyond line setting. All patients provided their written informed consent to the participation of clinical trials [8]. The dataset is limited to 41 patients who received an abdominal-chest CT discarding the other two who instead received MRI-based imaging. For each recruited patient, a chest-abdomen CT scan was performed for cancer disease staging. Each CT scan is complemented with the following clinical and personal history data (used in our learning model). Oncologists define the cancer extension through the *staging* step. The staging requires the use of imaging (usually CT-scan and PET) to characterize the level of disease spread in the subject body [191][188]. Especially in the advanced mUC stages, CT scans show multiple lesions, and radiologists/oncologists select the most significant ones (according to RECIST guidelines) for monitoring cancer evolution over time. The selection is carried out according to the RECIST guidelines that define inclusion criteria, CT scan procedure, patient assessment, lesion features, and how to monitor cancer over time. In particular, the enrolled subjects have been subdivided into:

- Patients showing a *Complete Response* (CR) to the medical treatment if all identified target lesions disappear at the end-treatment CT imaging;
- Patients showing a partial response (PR) to drug treatment if the target lesions are reduced by at least 30%;
- Patients showing a progressive disease (PD) if the target lesion increases by 20% of the *Longest Diameter* (LD);
- Patients showing *Stable Disease* (SD) if no significant increase or decrease is observed on the target lesions.

[8] Nr. D4191C00068 and MO29983, including the use of their clinical information for analysis approved by the Institutional review board (IRB) "Comitato Etico Catania 1", Catania, Italy

All the CT RECIST 1.1 compliant lesions have been collected for a total amount of 106 RECIST 1.1 findings. About 43 cases (target lesions) are referred to a complete/partial response or a disease stabilization following immunotherapy treatment (CR / PR / SD: Class 1), while 63 lesions are regarded to the disease progression despite anti-PD-L1 drug treatment (PD: Class 2). More details are reported on [190].

Data annotation, training procedure and evaluation metrics

An expert oncologist carried out manual data annotation through observing the whole CT scan and selecting all the target lesions according to RECIST 1.1 recommendation. Then, a 64 x 64 bounding box area (ROI) around each selected lesion over 16 consecutive slices has been extracted. Once the VOI has been selected in the first slice, our advanced software tool has automatically extracted the same VOI in the other slices (for a total of 16) to characterize the morphological, temporal dynamics of the lesion. As mentioned, CT data was complemented with 15 additional clinical and hematochemical data converted into numeric representation and suitably normalized for being processed by the proposed model. The VOIs were adequately labeled regarding the two classes previously identified and described (C1 and C2). However, the selection of the samples in each dataset split was not performed randomly, but in order to balance suitably the presence of the patients of the two considered classes (C1 includes patients with some response to immunotherapy, i.e., CR / PR / SD cases, while C2, patients with progressive disease (PD)), and consequently to ensure enough variability of the characteristics of the subjects. In particular, the dataset was configured as follows: 76 target lesions (28 of Class 1 and 48 of Class 2) were used for training and validation sessions, while the remaining 30 CT target image lesions (15 of Class 1 and 15 of Class 2) were used as the test set. To improve the validation reliability, we have implemented a cross-validation mechanism through k-fold. Specifically, we cross-validated our deep model by configuring k = 5 and reporting the results of this procedure in Table 3 (mean and standard deviation for the main performance indexes). The output of our deep model for each input data (16 x 64 x 64 VOI with additional clinical data) is a two class probability vector on which, during training, we compute negative log-likelihood loss with L_2 regularization weighted by a factor λ=0.0001. Mini-batch gradient descent was performed for minimizing the model loss, using the Adam optimizer, with an initial learning rate of 0.01 and a mini-batch size of 4. We perform random translation and rotation (with random degree value) along the spatial axis for data augmentation, consequently increasing the dataset dimension during the training session. Our deep model is implemented using the Pytorch framework. Experiments were carried out on a server with 2 Intel Xeon E5620 CPUs with four cores each, 96GB of RAM equipped with an Nvidia Quadro P6000 GPU with 24 Gbytes video memory.

Results and Discussion

We compared our architecture with such state-of-the-art deep architectures in order to provide performance benchmarks regarding the proposed application. Specifically, in order to evaluate the improvement in terms of performance compared to similar 2D and 3D deep networks, the authors have validated the performance of the used DenseNet backbone having the same architecture of our pipeline but without the inclusion of Self-Attention (Non-Local Blocks) mechanisms and separable

Model	Accuracy	Sensitivity	Specificity	F1-Score
2D ResNet-50	0.620 ± 0.052	0.604 ± 0.0078	0.636 ± 0.061	0.613 ± 0.058
3D DenseNet + H	0.713 ± 0.047	0.711 ± 0.041	0.716 ± 0.064	0.713 ± 0.043
3D DenseNet+SepConv+ H+ ResNet101	0.733 ± 0.049	0.729 ± 0.069	0.738 ± 0.047	0.731 ± 0.054
3D DenseNet	0.640 ± 0.034	0.636 ± 0.034	0.644 ± 0.048	0.638 ± 0.032
3D DenseNet+NLB+SepConv	0.878 ± 0.039	0.871 ± 0.054	0.884 ± 0.075	0.877 ± 0.041
Proposed	**0.922 ± 0.037**	**0.929 ± 0.053**	**0.916 ± 0.047**	**0.922 ± 0.038**
2D ResNet-101	0.829 ± 0.043	0.822 ± 0.054	0.836 ± 0.061	0.828 ± 0.043
3D DenseNet-201	0.856 ± 0.033	0.871 ± 0.047	0.840 ± 0.055	0.858 ± 0.032
2D VGG-19	0.667 ± 0.033	0.662 ± 0.069	0.671 ± 0.059	0.664 ± 0.041
Previous method [188]	0.861 ± 0.023	0.815 ± 0.011	0.883 ± 0.048	0.810 ± 0.037

Table 3.3: Experimental performance benchmarking (mean ± standard deviation).

convolutions. In addition, the performance of the proposed method has been compared with respect to such classic architectures: ResNet-50, ResNet-101, VGG-19 and 3D extension of the classical DenseNet-201. The Table 3.3 reports the collected experimental results and comparions. The implemented 3D DenseNet backbone baseline (3D DenseNet) showed an accuracy of 0.640 ± 0.034 significantly lower than our full pipeline which shows higher accuracy equal to 0.922 ±0.037. Also, in terms of sensitivity, specificity and F1-score, our architecture is significantly more performer (0.929 ± 0.053 , 0.916 ± 0.047 and 0.922 ± 0.038 respectively) than the simple DenseNet backbone thus confirming the improvements that can be obtained in particular through the use of self-attention techniques and separable layers. Specifically, feature maps that suitably weight the spatiotemporal dependencies of the CT imaging selected target lesion (i.e., the result of non-local blocks application) provide more discriminative features to the FC stack. Moreover, the joint contribution of non-local blocks and separable convolution layers allows to generate feature maps having an informative content that best characterizes the spatiotemporal dependencies between the CT imaging VOIs and treatment response of the associated patient. The experiments revealed a considerable reduction in the overall performance of the tested deep networks if as inputs we only use CT visual imaging and do not integrate with blood and medical history. More in details, the proposed architecture with visual input but without hematochemical data dropped in performance as it showed 0.878 ± 0.039 (Accuracy) , 0.871 ± 0.054 (Sensitivity), 0.884 ± 0.075 (Specificity) and 0.877 ± 0.075 (F1-score) significantly lower with respect to the same proposed pipeline with hematochemical data. The performance of our architecture is significantly higher than the compared 3D deep classifier, as reported in Table 3.3. This confirms that the Self-Attention mechanisms realized through the inclusion of non-local blocks with embedded gaussian setup together with the separable layers allow to obtain more discriminative and representative features maps (with respect to deeper network as the 3D-DenseNet-201) of the correlation with the response to immunotherapy treatment.

Our proposed deep architecture aims not only to offer valid medical assistance to the physician but rather to highlight the most predictive visual patterns. In doing so, we have attempted to investigate and adopt one of the most promising self-deep features of explanatory methods already introduced in the scientific literature. The authors propose the usage of GradCAM introduced by Selvaraju et al. [187]. The GradCAM approach uses the gradient for the generated convolutional features as a

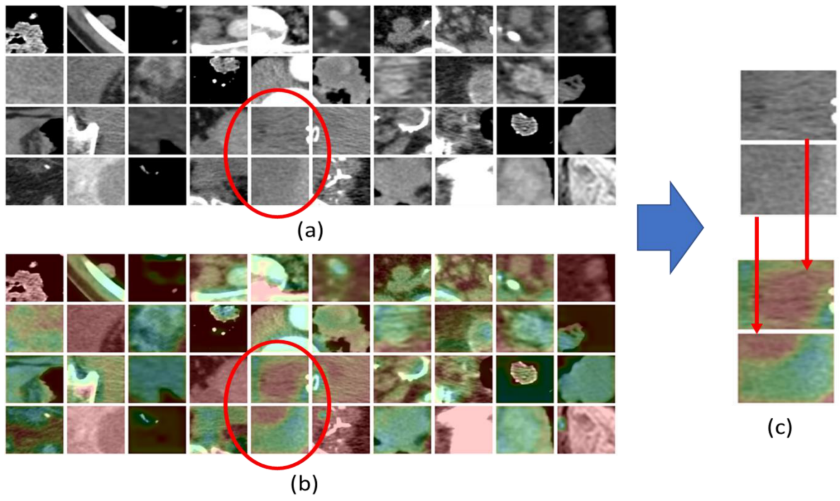


Figure 3.12: (a) RECIST (Response Evaluation Criteria in Solid Tumors) 1.1 compliant CT target lesions. (b) The corresponding Grad-CAM generated saliency maps. (c) A detail of the salient part of the processed RECIST lesion.

classification score to understand which parts of the input image are most significant for classification. Using GradCAM, we attempted to understand which parts of the ROI extracted from the chest-abdomen CT images (containing the target RECIST compliant lesion) were more significant for our 3D pipeline processing. As evident from Fig. 3.6.1, for some processed lesions (ROI), the GradCAM analysis highlighted such areas with greater salience (red area) than the others (green area). The salient visual areas of such input ROI-lesion are those that most contribute to the performance of the deep network, i.e., those that are best represented in the feature maps. In [190], we reported more discussion related to our proposed method and collected dataset as well as the visual representation of the most significant features of ROIs by using GRAD-CAM.

3.6.2 Approach based on Non-linear Generative Model

Similar to the previous study [190], we propose an innovative approach based on the use of a non-linear cellular architecture with a deep downstream classifier for selecting and properly augmenting 2D features from chest-abdomen CT images toward improving immunotherapy outcome prediction [192].

Immunotherapy is currently considered the last frontier of cancer treatment [193], [194]. Cancer cells are normally recognized by the immune system, which triggers an attack by T lymphocytes. However, this body defense mechanism is not always effective because cancer cells are able to implement a whole series of escape strategies. One of these benefits from the immune system's self-regulation mechanism through a series of proteins that act as "accelerators" or "brakes" on T cells [193]–[196]. A promising immunotherapy strategy is based on the inhibition of ICIs, i.e., on the use of specific antibodies to re-enable the immune system (previously disabled by the cancer cells) and thus increasing the ability of T lymphocytes to deal with tumors [193]–[195]. In this contribution, we will focus on ICIs-based immunotherapy treatments that act on the PD-1 receptor [195]. Scientific studies have shown that cancer cells "defend" themselves from T lymphocytes using a molecule present on their membrane, called PD-L1, which binds to the lymphocyte PD-1 receptor to

disable the protective action of the T lymphocytes. The ICIs anti PD-1 / PD-L1 immunotherapy treatment aims to inhibit the action of PD-1 / PD-L1 receptors that prevent T lymphocytes from recognizing and destroying cancer cells [195], [196]. In this regard, we propose an innovative and less invasive pipeline based on a DL algorithm for classifying visual features extracted from radiological images (chest-abdomen CT-scan) of patients with a bladder cancer diagnosis. Through applying this pipeline and using a specifically configured Cellular Non-Linear (or Neural) Network (CNNs), CT cancer lesions are properly identified and augmented.

The proposed Deep Network Framework

First of all, the proposed semi-automatic pipeline performs a bounding box segmentation of the CT lesion identified by the experienced physicians, which defines the ROI to which the proposed predictive pipeline is applied. Those ROI images have different biologically visual features as they have been extracted from various body-sites of mUC metastasis (lungs, abdominal organs, bladder, lymph nodes, etc .). The segmented ROIs (CT lesions) are further processed by ad-hoc configured and extended 2D Cellular Non-Linear Network (2D-CNN), which generates a series of augmented domain-agnostic features to be properly classified by downstream 2D Deep classifier (2D-DNN). The designed 2D-DNN discriminates augmented visual features for predicting correlated patients who potentially show some response to immunotherapy treatment (CR: Complete response / PR: Partial response /SD: Stable disease) by those who instead have a progression of the disease (PD: progressive disease).

The Bounding Box Segmentation block. This block performs a semi-automatic CT-scan lesion segmentation driven by experienced oncologists/radiologists. Specifically, starting from the whole chest-abdomen CT scan of each patient, the oncologist/radiologist manually selected a target characteristic cancerous lesion according to RECIST revision 1.1 criteria [189]. Similar to previous research study [190], an imaging software was used to automatically select ROI according to certain spatial, dimensional, or morphological criteria. After performing this selection, a MxN bounding box ROI around (centered), the lesion is automatically extracted. The dimensional setup (M, N) can range, and it does not affect the overall pipeline performance as the segmented lesion will be further re-scaled according to the input size of the downstream 2D-DNN classifier.

The 2D-CNN Features Generative Model. This block performs further processing of the collected ROIs in order to leverage discriminating visual features. The proposed generative model is based on a properly configured and extended version of 2D Cellular Non-linear Networks (CNNs) [197]. The CNN can be defined as high speed local interconnected computing array of analog processors called "cells" [197]. The basic unit of CNN is the cell. The CNN processing is configured through the instructions provided by the so-called cloning templates [197]. Each cell of the CNN may be considered a dynamical system arranged into a topological 2D or 3D structure. The CNN cells interact with each other within its neighborhood configured by a heuristically ad-hoc defined radius [197]. Specifically, the 2D-CNN is a *transient-response* CNN. The transformation of the input data is performed at the transient stage, where every single cell of the CNN dynamically evolves from the initial state along the trajectory that converges to the CNN steady-state [197], [198]. In this way, we are able to extract intermediate 2D-CNN transformation features of the input, which will be used as augmented-generated features. More details on the proposed Generative model can be found on [192].

The 2D-DNN Classifier with Decision System. This block aims to learn the 2D-CNN generated augmented features to properly classify the patients eligible for immunotherapy treatment. Different 2D deep classifiers were investigated to test the computational complexity over the implemented embedded hardware Point of Care and determine those that provide the best predictive performance. Specifically, the following 2D deep classifier backbones have been validated: ResNet-18, VGG-19, Xception, MobileNetV2, GoogleNet, AlexNet. Finally, a NasNetMobile based 2D Deep Classifier has been tested [199]. For each patient, the RECIST 1.1 compliant target CT lesions (ROIs) are selected by the introduced semi-automatic segmentation block. The re-scaled ROIs are augmented through the proposed 2D-CNN generative engine. The generated augmented features maps will then be classified by the 2D-DNN, which provides a probability estimation of belonging to class 1 (CR / PR: complete or partial response to immunotherapy treatment or SD: stable disease) or class 2 (PD: progressive disease). We also designed a Decision System block. The task of the Decision Systems for each patient is to determine the main 2D-DNN classification rate of the CNNs generated features. This predominant classification will be the definitive classification of the patient. The output class of the Decision Systems becomes the definitive classification (immunotherapy outcome prediction) related to the analyzed patient.

Dataset

We retrospectively analysed a dataset of 106 mUC cancer lesions (cases) extracted from the chest-abdomen CT scans of trial-recruited patients with histologically confirmed bladder cancer. Patients with multiple non-overlapped RECIST-compliant lesions have been analysed for each target lesion detected on CT examination. The recruited patients had histologically confirmed bladder cancer (mUC) progressing after platinum-based chemotherapy and were treated with a PD-L1 ICIs immunotherapy agent in the second-line setting. Some statistical information about the recruited dataset is reported on [192]. The lesions thus collected in both training and testing were then augmented through the proposed generative model based on 2D-CNN. Therefore, considering that for each CT image lesion a generative engine of $m = 97$ CNN models setup (3x3 cloning templates and 1x1 biases) has been defined, the augmented dataset will be composed of a total of $7,372$ (76×97) image features ($2,716$ images of Class 1; $4,656$ images of Class 2) for the training and validation set and $2,910$ image features ($1,455$ images for each class) for the test set.

Results and Discussion

Table 3.4 represents the results in terms of accuracy, sensitivity, and specificity, using each selected 2D-DNN classifier with a CT image dataset augmented through the proposed 2D-CNN generative model. Although some architectures show 100% in terms of sensitivity, the deep architectures showed acceptable performance even in specificity are preferred. An interesting trade-off in classification performance (accuracy, sensitivity, specificity) was obtained by ResNet-18 (Accuracy: 93.33% Sensitivity: 93.33% and Specificity 93.33%) and VGG-19 (Accuracy: 93.33% Sensitivity: 100.00% and Specificity 86.66%). The contribution of the 2D-CNN generative model is confirmed through the benchmarks reported in Table 3.5. We report the comparison results of the most performer architectures reported in Table 3.4 (ResNet-18 and VGG-19) but without 2D-CNN generative model and by using classic input augmentation approaches [201]. The results in terms of overall performance were

2D-Deep Classifier Backbone	Metrics			
	Accuracy	*Sensitivity*	*Specificity*	*Input Size*
ResNet-18	93.33%	93.33%	93.33%	224x224
VGG-19	93.33%	100.00%	86.66%	224x224
XCeption	86.66%	93.33%	80.00%	299x299
MobileNetV2	83.33%	100.00%	66.66%	224x224
GoogleNet	86.66%	86.66%	86.66%	224x224
AlexNet	83.33%	80.00%	86.66%	227x227
NasNetMobile	76.66%	80.00%	73.33%	224x224
Previous method [200]	86.05%	80.00%	89.29%	40x40

Table 3.4: 2D-DNN Performance Benchmark - 2D-CNN Dataset Augmentation Model

Deep Classifier Backbone	Metrics		
	Accuracy	*Sensitivity*	*Specificity*
ResNet-18	86, 66%	100.00%	78.66%
VGG-19	73.33%	86.66%	60.00%
3D-DenseNet	83.33%	86.66%	80.00%
ResNet-101	76.66%	80.00%	73.33%

Table 3.5: 2D-DNN Performance Benchmark - Classical Dataset Augmentation Method

significantly lower than the same architectures trained with a dataset augmented through 2D-CNN. Although ResNet-18 grew in sensitivity compared to the same architecture with the 2D-CNN generative model (100% compared to 93.33%), it considerably under-performed in specificity (78.66% compared to 93.33%), thus limiting the overall performance of the pipeline. VGG-19 also showed significantly lower performance than the same with the upstream 2D-CNN generative model. As confirmed by experimental results reported in Table 3.4 and 3.5, the proposed pipeline shows very promising performances both in terms of accuracy, sensitivity and specificity. The improved performance stems from the combination of a deep high-capability classifier and 2D-CNN augmented training set [78]. The proposed 2D-CNN augmentation pipeline enables the downstream 2D-DNN to perform a greater and deeper exploration of the segmented RECIST 1.1 chest-abdomen CT image-lesions identifying more discriminative visual patterns in an originally limited clinical dataset. The proposed pipeline has been validated overall recruited RECIST 1.1 compliant lesions. This confirms that the method is not strongly affected by the lesion selection performed by the radiologist.

3.7 Future Works

In this section, we report some approaches we aim to further investigate in our future research activity in the context of Medical Imaging. In discussing our future directions, we provide insights related to what the scientific community is focusing on. Among several avenues considered to undertake the implementation of advanced DL solutions, the researchers include:

Transfer Learning. In several applications, ranging from biological to medical, training a deep model from scratch (i.e., starting from randomly initialized weights) represents a challenging task, especially whether the dataset includes a small number of samples. In order to overcome this issue, the scientific community has proposed the *transfer learning* strategy. Generally, transfer learning is performed by training the network using the weights referred to another network already trained on a larger dataset (e.g., ImageNet). The advantage of this procedure consists in optimizing the training process, obtaining effective results rapidly. Due to its effectiveness, Transfer Learning has prominently been applied in the medical imaging area over the last few years, enhancing the DL performance on limited dataset [202], [203].

Generating synthetic data. The scientific community is also investigating the application of Generative Adversarial Networks (GANs) [204] in the medical field. The scarce availability of large datasets could be addressed by generating synthetic data. In this regard, the most reliable approach to perform this task is based on the recently introduced Generative Adversarial Networks (GAN). Briefly, GAN networks are characterized by two components: the generator and the descriptor, two separate neural networks devoted to accomplishing different purposes. In particular, the two neural networks "compete" against each other. The generator receives random noise as input to generate new data as similar as possible to the real ones, while the discriminator learns how to distinguish the new synthetic data received as input from the real data. The discriminator implements a binary classification to define whether images are real or fake, computing the final probability. The two networks cease to compete when the descriptor reaches a probability of 0.5, which means that the nature of data is indistinguishable. In the medical context, the use of GAN allows generating synthetic data referring to multiple patients, maintaining a certain inter-patient and intra-patient variability, and augmenting training data.

Unsupervised learning. The scarce availability of labeled datasets has hindered the application of DL algorithms in a wide range of Medical Imaging problems. In this regard, the scientific community has actively searched for the most effective approaches to address this issue. The recent trend in the medical domain is to use unsupervised learning methods due to their ability to exploit unlabeled data, grouping entities according to their similarity. In the medical field, these approaches provide a more robust performance without being fooled by biased knowledge, which affects its supervised counterpart. Moreover, the unsupervised approaches provided impressive performance with lower cost as they do not require a substantial manual effort to create labels, leading to time savings [205].

3.8 Conclusions

In this chapter, we have investigated medical imaging applications from different perspectives, discussing in detail our research projects and the challenges encountered in this domain. Specifically, we listed the main issues related to applying DL algorithms in a clinical context, focusing on the issues concerning the non-availability of high-quality datasets. Since DL methods inspire their potentials by processing a massive amount of data, large datasets have become a crucial point to building a powerful tool. However, the main barrier to collecting quality data in this field is the labeling process usually performed by a human expert, leading to a time-consuming

and error-prone process. Our approaches were developed with the intention to furnish advanced solutions to overcome the mentioned difficulties for modality classification images and other medical imaging applications, such as the detection of COVID-19 disease and the prediction of immunotherapy outcomes. Note that the achieved results are preliminary. As mentioned in the previous section, we aim to further investigate more reliable methods for dealing with limited datasets, which represents the main drawback of these kinds of applications. Although DL techniques are revolutionizing the medical field, the skepticism towards DL methods, derived by non-comprehensive knowledge, has hindered the spread of such technologies in this challenging area. For this reason, we also treated the problem of *interpretability* of DL models, which is a crucial facet in the medical context in order to make clinicians understand what underpin them. Despite these limitations, we would expect that the application of DL techniques will affect the future of the medical domain, promoting the synergy between clinicians and highly skilled professionals in the AI domain (i.e., computer scientists, engineers, mathematicians) and a cultural change in which healthcare professionals are proactively involved in the designing of effective AI approaches for medical tasks.

Chapter 4

Findings, Limitations and Perspectives

DL approaches have demonstrated their efficiency and efficacy on several fundamental data processing tasks. Hence, data analysis plays an important role in complex scenarios. Particularly in automotive and medical imaging, where the rapid outcome predictions are gaining crucial importance, a DL technique devoted to automatically learning the most descriptive and salient features from a high volume of data can dramatically lead to the accomplishment of challenging tasks. DL methods enhance the investigation of complex non-linear data and reduce the involvement of human expert analysis. However, these techniques often suffer from overfitting problems or fail to extract enough information when data are limited as to their performance directly depends on the amount and quality of available data. The broad scope of this dissertation is to propose advanced DL pipelines in the context of automotive and medical imaging, and on that basis, to outline which challenges we encountered during their development and to which extent it is possible to implement solutions to overcome them.

In Chapter 2 we reported our research works in the automotive field, focusing on methods that used physiological signals to assess the psychophysical status of a car driver. An extensive amount of studies have demonstrated that the analysis of such data underpins an effective pre-emergency warning system for the driver, avoiding accidents due to an inadequate condition (e.g., altered state due to fatigue, etc.). After providing a comprehensive discussion on the current ADAS functions, we outlined the proposed DL architectures and the challenges encountered in this research field. In doing so, our research highlighted that:

- The current ADAS functions are devoted to guaranteeing the driver's safety. In this regard, researchers have spent many efforts to design several AI algorithms which process data derived from different sources (especially automotive-compliant devices). In this regard, the scientific community exploits advanced CV and DL algorithms to process data based on vehicular, behavioral, and physiological parameters.
- Since the acquisition of physiological signals, such as the PPG, may introduce motion artifacts or noise, it is required to define ad-hoc pre-processing pipelines in order to generate a denoised signal.
- DL approaches require a computational workload that appropriate tools must manage. In this regard, the scientific community in the automotive industry is still investigating the most efficient ways to integrate automotive-compliant hardware tools running both CV and DL algorithms.

These findings furnished motivation to design the following approaches:

- We designated promising DL pipelines to process both subject's physiological signals and blood pressure measurements in order to build an advanced warning system that prevents vehicles accidents through analyzing his/her drowsiness. The primary purpose is to detect driver fatigue and inattention in the real-time driving scenario in a more reliable way.
- We implemented approaches that extract information about driver status by acquiring visual data (i.e., facial landmarks) to provide insights about the vigilance level of a subject without involving the use of sensors. Since the latter must follow strict protocols to be integrated into ADAS systems, installing these devices in the automotive environment is not often feasible.
- We proposed an ad-hoc filtering pipeline to remove noise and artifacts from the PPG signals to enhance its processing by DL algorithms.
- Automotive-grade devices' hardware resources are limited, with respect to the once usually required to implement modern DL and CV algorithms for detecting driver drowsiness. We compared the performances of a laptop and an existing commercial automotive-grade device developed by STMicroelectronics in running a CV algorithm. Although our results show that the computation time is still quite long to be used in a real-time scenario, we demonstrated that the used board successfully runs advanced algorithms, encouraging us to pursue further studies in this direction. Hence, our future research will focus on running DL algorithms on this board.

Chapter 3 presents the research works done in the field of medical imaging. In this instance, we listed the drawbacks of the medical domain, and designed approaches covering three main topics:

- The interpretability of a DL model. In the medical domain, achieving impressive results in terms of accuracy is not a sufficient achievement. Due to the ambiguous nature of deep networks, their predictions are not clear for experts without an IT background, such as clinicians. Hence, it is required to evaluate how much a model is interpretable. In Chapter 3, we provided a definition of interpretability as well as solutions proposed by the scientific community to improve this crucial property.
- The analysis of multi-modal data. Medical Imaging refers to the analysis of data derived from multiple medical examinations. Depending on the examination, the medical manufacturer produces images with a specific modality. As a result, medical image datasets present a massive amount of data with different modalities. In this context, defining a DL algorithm for automatic indexing and retrieval of medical images may provide an effective tool to assist clinicians in defining diagnoses and treatments for patients.
- Non-availability of labeled data. The lack of labeled datasets is a crucial point in the medical imaging field. As discussed in Chapter 3, the labeling step represents a time-consuming process, as it often requires the supervision of a high-skilled clinician. This also entails an expensive cost in terms of human resources. Hence, unlabeled data are not considered adequate as input for deep networks, which are affected by the quality of data used to perform the training process. Despite the recent attempts to develop unsupervised learning

methods for automatically grouping medical data into n classes, the analysis of medical images performed by a qualified expert is still predominant in the medical imaging domain.

In this context, our research has led to the definition of:

- *Methods for evaluating interpretability of a DL model*. As discussed in Chapter 3, current literature demonstrated the strong correlation between the robustness and interpretability of a DL model. In outlining a DL method to identify patients who could be gene-addicted or not, we also improved the model robustness to adversarial attacks (small alterations added to the training images) by applying a defense technique known as *adversarial training*. Finally, we evaluated the robustness by using the Frobenius norm of the Jacobian matrix. The results showed that *adversarial training* leads to more interpretable gradients and enhances the interpretability of the DL process.

- *DL algorithms for multi-modal data classification*. We used pre-trained networks, VGGNet16 and ResNet101, to define a baseline for the modality classification of medical images. In this regard, we compared our results with the ones obtained by NFNets, an advanced DL architecture recently proposed by Google's DeepMind group. The obtained results have confirmed that deep networks are able to classify with impressive accuracy images having different modalities, thus defining an effective tool to distinguish medical data.

- *DL approaches for disease detection and outcome predictions*. We adopted advanced DL techniques for the detection of COVID-19 disease. In this regard, we applied an existing approach based on the Inf-Net neural networks to perform disease segmentation on chest CT scans. We also defined DL architecture based on Non-Local neural networks to establish whether a patient responds to immunotherapy treatment or not. Although our contribution is limited to collecting and analyzing data, we detail these research works as they provide a significant progress in the context of immunotherapy treatment outcomes prediction.

In this dissertation, we have investigated several aspects related to DL approaches in the fields of automotive and medical imaging. More broadly, our research activity revealed common issues in both research areas regarding the non-availability of high-quality datasets. This issue have always plagued the AI domain, hampering the spread of DL approaches in challenging contexts. Despite facing with several issues, we attempted to design innovative pipelines to overcome the limited samples The results related to the automotive applications confirmed that the proposed solutions encouraging more in-depth analysis, whereas the reported outcomes in medical imaging refer to preliminary findings which need to be further investigated.

4.1 Future Works

In the context of the automotive field, future works can consider combining physiological signals with driving scenario information (e.g., road surface segmentation) to develop safety systems for reducing the number of automobile accidents. Moreover, recent methods exploiting thermal camera devices to collect the car driver's visual information under low-light conditions could also be taken into account [206]. With regard to the medical imaging domain, future works can be devoted to using transfer

learning approaches to perform new tasks. Furthermore, the generation of synthetic data through the use of GAN models [207] can be considered. Finally, unsupervised learning can be developed to cope with the absence of labeled data from the medical field.

Appendix A

Deep Learning for Pollen Grain Microscope Images

In this section, we summarized the research activity carried out thanks to the collaboration with Ferrero HCo, which financed the project and allowed the collection of aerobiological samples from hazelnut plantations. The described activity consists in the definition of a large-scale dataset of microscope pollen grain images, collected from aerobiological samples, including four species of pollen grains and an additional class of objects that could often be misclassified as pollen (e.g., air bubbles, dust, etc.). More than 13,000 objects have been detected from microscope images and labeled by experts. Moreover, ad-hoc pipelines to segment pollen grains from images are also proposed to perform pollen classification by using DL approaches in order to assist scientists in grouping pollen objects by species.

A.1 Detection and classification of pollen grain microscope images

A.1.1 Background and Motivations

With the rapid development of technologies in the field of AI, image data analysis has attracted much research attention over the last few years. In particular, typical problems in CV and DL field are related to image classification. Indeed, image classification embraces several issues such as discriminative feature extraction. The rapid emergence in developing innovative pipeline to solve image classification has led to the spread of AI methods for extracting features from images. In particular, one of the application areas that benefits from the power of advanced image classification technologies is aerobiology. Aerobiology, the discipline that studies airborne biological particles and their dispersal mechanisms, has a crucial role in several fields such as medicine, biology, and agronomy, with direct and non-direct effects on the economy and public health. Estimating the abundance of airborne allergenic pollen and fungal spores allows us to evaluate the associated health risk and the potential infectious diseases [208] on both humans [209] and plants [210] for certain periods. The amount of airborne pollen can be considered as a proxy to plant phenology and flowering intensity, thus leading to its integration in many yield forecasting systems applied to commercially important crops [211], [212]. Despite its effectiveness, the involvement of an expert to analyze images in microscopy is a time-consuming task that has hindered the application of aerobiology to those and new sectors [213]. Despite of the various efforts to develop devices that allow the identification and classification of pollen grains without the need of end-user intervention [214], [215], the observation and discrimination of features from relevant entities performed by

(a) (b) (c) (d) (e)

Figure A.1: Examples of acquired samples. (a) *Corylus avellana* (well-developed pollen grains), (b) *Corylus avellana* (anomalous pollen grains), (c) *Alnus* (well-developed pollen grains), (d) Debris, (e) *Cupressaceae*.

qualified experts is still predominant [216]. For the above reasons, research on methods for automatic classification of pollen grains will foster the development of tools for aerobiologists that could be used without the need for special equipment other than a common bright-field microscope, a camera for image acquisition and a computer for data processing. In this regard, DL approaches provided a remarkable contribution in the field of aerobiology. One of the key elements of DL success is the availability of a high amount of annotated data. Indeed, the performances of the DL techniques scale with the amount of data, promoting the definition of large-scale datasets in different scientific fields.

A.1.2 Dataset

The main contribution of this research activity is the availability of a novel dataset that includes 13,416 objects of pollen grains. Experts in aerobiology manually labeled each object by using a web-based tool. In order to pre-process and analyze the aerobiological samples, they used segments of tape in which the pollen grains adhered. Each segment of tape was placed on a rotating drum, moved at 2 mm h-1 under a suction hole. The daily segments of the pollen grains have been inspected by making use of a Leitz Diaplan bright-field microscope and a 5 MP CMOS sensor. By following the standard procedures, the pollen walls placed on the microscope slides were selectively stained with a mounting medium containing basic fuchsin (0.08 % gelatin, 0.44% glycerin, 0.015% liquefied phenol, 0.0015% basic fuchsin in aqueous solution). The dataset is composed of more than 13,000 objects spanning over 5 different categories: (1) *Corylus Avellana* (well-developed pollen grains), (2) *Corylus Avellana* (anomalous pollen grains), (3) Alnus, (4) Debris, (5) *Cupressaceae* (see Fig. A.1). We collected objects for each class alongside the related binary mask and the segmented object with green background. An amount of 63 images in the dataset represents pollen grain objects overlapped with a non-pollen one. To avoid the misclassification of these images, we decided not to insert them into the dataset used for the experiments. Considering the small number of observations related to *Cupressaceae* class (43), we did not include them in the dataset used for the experiments. Therefore, the total number of objects in the dataset for experiments is 13,310 (see Fig. A.2).

A.1.3 Proposed solutions

The Processing pipeline

This section introduces the designed pipeline which consists of three main stages: pre-processing, segmentation and mask post-processing.

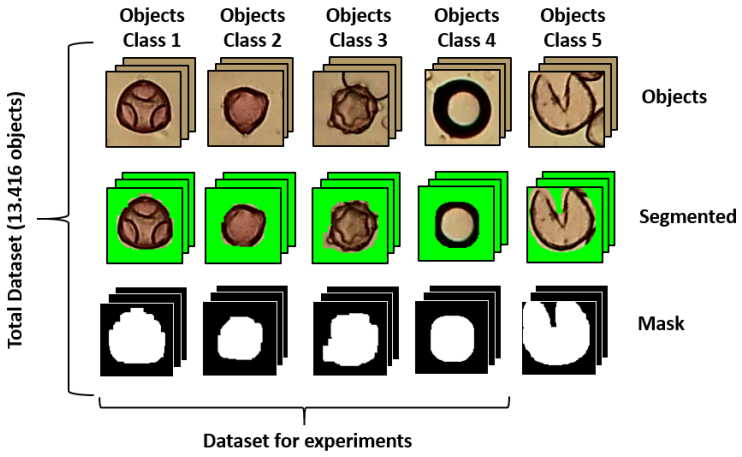


Figure A.2: The proposed dataset that includes for each object the related binary mask and the segmented object.

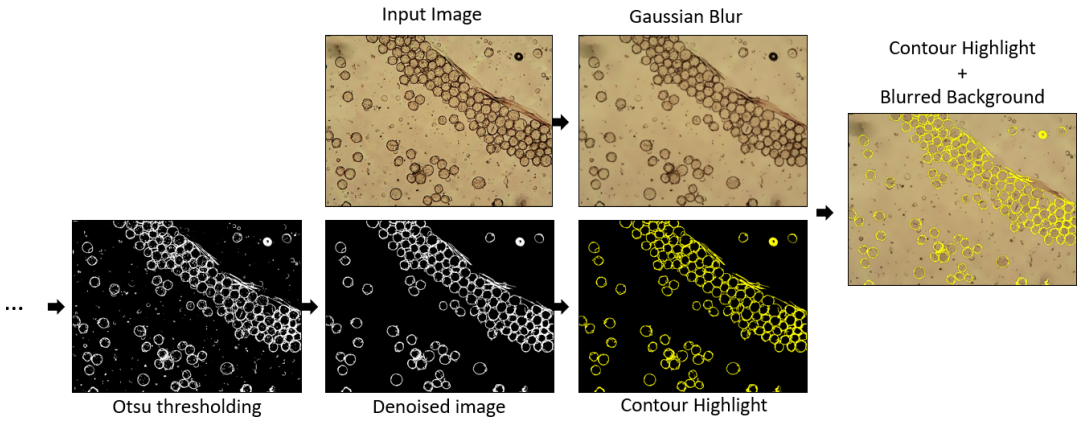


Figure A.3: Pipeline used to highlight contours objects.

Pre-processing. As previously mentioned, the automatic detection of pollen grains is a challenging task due to the presence of heavy background noise in the digitalized slides scans, which can drastically affect the performance of segmentation methods. For this reason, we designed a proper pre-processing pipeline with the aim of reducing background noise and improving the high segmentation quality in the presence of several problematic conditions deriving either from the manual sectioning of the aerobiological sample itself (i.e., debris and dust and fungal spores) or from the mounting technique (air bubbles). The first step of the proposed pipeline consists of applying the OpenCV implementation of the mean shift algorithm [1]. The output of this procedure is an image with color gradients, and fine-grain texture flattened. In the second stage, we split the pipeline into two different steps. First, we developed a procedure to smooth background artifacts to improve object detection in the foreground. In order to discriminate the foreground from the background of the mean-shifted image, we used Otsu's method [217] in combination with a binary threshold, set to 127. Under the supervision of experts in aerobiology, we observed that pollen objects are usually higher than 500 in diameter. Based on these assumptions, we removed all the objects with a size smaller than this by using connected

[1] `https://docs.opencv.org/2.4/modules/imgproc/doc/filtering.html?highlight=meanshiftfiltering`

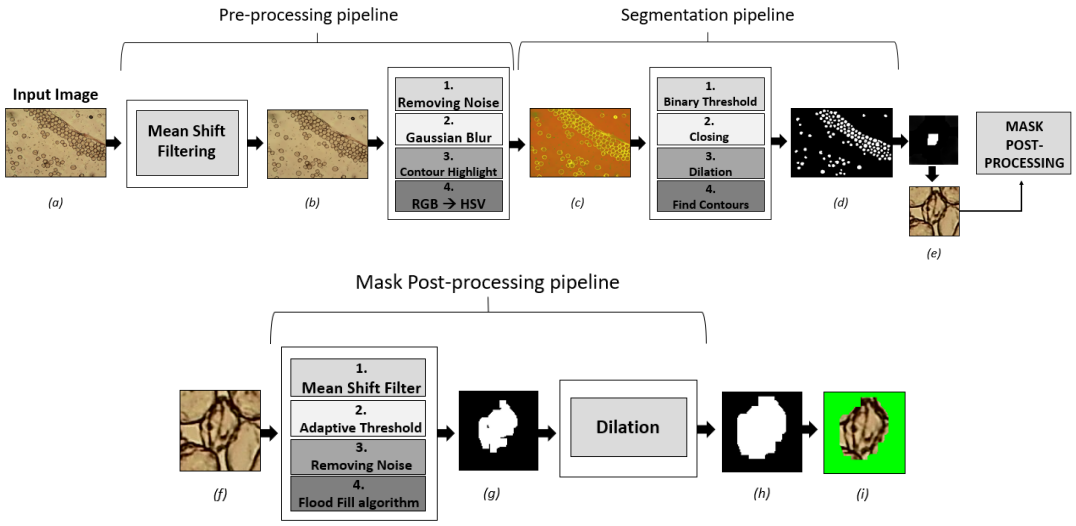


Figure A.4: The overall pipeline. (a) Image of an aerobiological sample, (b) image after applying mean shift filtering function, (c) the resulting image after converting color space from RGB to HSV derived from previous steps, (d) the mask generated by applying binary threshold, closing and dilate operators, (e) image after applying binary mask to input image (f) the detected object, (g) the resulting binary mask after applying a mean shift filter and adaptive threshold, (h) the obtained binary image and (i) the related segmented object from original patch.

components with eight neighbors. Finally, we applied the resulting binary mask to the input image and changed the color of the detect contours in yellow. The second step of this stage consists of applying an 11×11 Gaussian kernel filter to the input image to blur the background artifacts. In the final stage, we combined both the output images from previous steps. This step is essential not only to distinguish objects of the foreground from the other ones in the background but also to guarantee that the following image processing will not damage the object contours, influencing the effectiveness of the image segmentation pipeline. In Fig. A.3, we illustrated the described steps to highlight objects' contours.

Segmentation pipeline. The object segmentation pipeline aims to partition an image by maximizing the pollen grains detection alongside reducing the detection of non-pollen ones (i.e., dust, artifacts, etc.). With this aim, after changing the color space of the output image of the previous stage from RGB to HSV, we transformed the image in grayscale and applied a binary threshold to it. In order to reduce the noise background generated by previous image processing steps, we applied a closing operator followed by dilation using a 3×3 kernel for both operations. In this instance, we implemented the flood fill algorithm intending to distinguish the foreground from the background by reassigning the values of all neighboring pixels of a given point with a required uniform color. Therefore, all the objects of interest have been filled with black color, whereas the background has been filled with white color. Through analyzing connected components in the image, we also removed objects with a size smaller than 100 pixels in diameter.

Mask post-processing pipeline. The developed procedure for mask post-processing is fundamental to improve the image quality of the output image derived from the previous steps. The main idea is to generate a binary mask for each object, developing a closed outline for the object's boundary in a manner that outlines the object

contours tightly and minimizes background noise. To this aim, we applied a pool of effective image processing functions. After applying a mean shift algorithm, we used an adaptive threshold function. Formally, the function converts a grayscale image into a binary image by using a given threshold. The threshold value is calculated individually for each pixel in the input image. Specifically, we used an adaptive Gaussian threshold. In the final step, a procedure based on selecting connected components (objects) with a size greater than 150 pixels of diameter is developed in order to reduce background noise. Once applied the flood fill algorithm, we used a dilation operator with a 3×3 kernel full of ones to increase the object's size in binary mask image, iterating the process for 5 times. In Fig. A.4, we reported the overall pipeline.

A.1.4 Experiments

This section presents the results obtained for a pool of Machine Learning algorithms, comparing the performance of the techniques herein proposed for pollen grains classification. Also, we addressed the problem of the imbalanced dataset and how we deal with it.

Experiments with LBP and HOG features

In this stage, we carried out our experiments via classical Machine Learning techniques, such as Random Forest, Support Vector Machines (SVM), AdaBoost, among others [218]. In order to introduce the data to a classifier, we first needed to extract features from the raw data. Specifically, we investigated object texture by using Local Binary Pattern (LBP) [219] and Histogram of Oriented Gradient (HOG) [220]. In order to perform classification, we divided the data into a training set and a test set. In doing so, we have considered the first 85% of the data as the training set and the remaining 15% as the test set. In this stage, the classification is performed with the dataset composed of segmented images with green background in order to consider the only texture of the detected object. Also, we carried out our experiments by using the following models: Linear Support Vector Machine (SVM), RBF SVM, Random Forest, AdaBoost, Multi-Layer Perceptron (MLP) [218]. To handle the problem of imbalanced data, we first selected optimal hyperparameters by applying the Grid Search algorithm [2] and computing the average accuracy at each run. Therefore, we evaluated the performance of each classifier by using the proposed imbalanced dataset, relying on optimal hyperparameters that have been selected from the previous step. To carry out our experiments, we used penalized classification models for SVM and Random Forest algorithms. Penalized models help focus on the minority class by adjusting weights inversely proportional to class frequencies in the training data. The major issue of the imbalanced class dataset is the lack of samples of a given class, which could lead to poor results. Also, we performed a stratified train-test splitting to ensure that both training and test set featuring the same percentage of classes.

To evaluate the performance of each classifier, we used the weighted F1 score for quantitative evaluation [221], which represents a more reliable performance metric than accuracy in this context. The weighted F1 score function calculates the F1 metrics for each class and its average weighted by support (i.e., the number of true

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

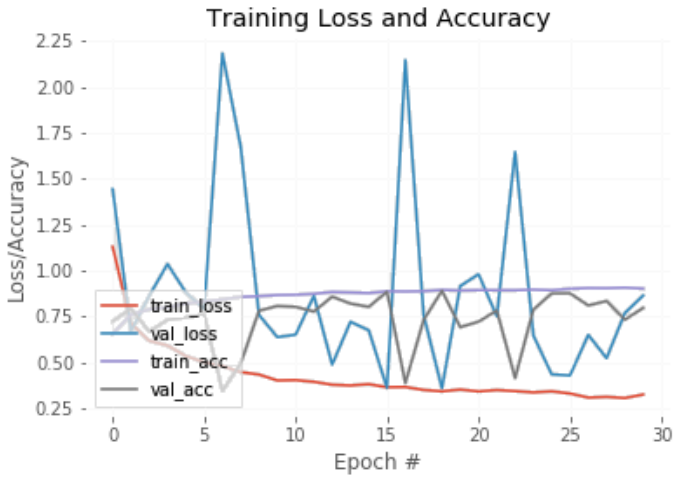
Methods	Parameters		HOG Accuracy	HOG F1-score
LINEAR SVM	$C = 1000$		0.7646	0.7673
RBF SVM	G = 0.1	C = 1000	**0.8658**	**0.8566**
RANDOM FOREST	$EST = 10$		0.7616	0.7124
ADABOOST	LR = 0.5	EST = 500	0.7752	0.7627
MLP	a = 0.1	EST = 300	0.8493	0.8431
Methods	**Parameters**		**LBP Accuracy**	**LBP F1-score**
LINEAR SVM	$C = 100$		0.7446	0.7439
RBF SVM	G = 1.0	C = 1000	0.6430	0.6714
RANDOM FOREST	$EST = 1000$		0.7792	0.7387
ADABOOST	LR = 1.0	EST = 100	0.7722	0.7487
MLP	a = 0.0001	EST = 500	0.8002	0.7764

Table A.1: Comparison between the best results by using HOG and LBP features.

instances for each class). With regard to LBP, Table A.1 reports the experiments related to the LBP feature representation for the input pollen images, considering the evaluated parameters for each employed classification algorithm. Results show that MLP leads to the best results in terms of accuracy (0.8002) and F1 score (0.7764) with an alpha value equal to 0.0001 and a number of estimators equal to 500. We also observed that Linear SVM, Random Forest, and AdaBoost performed accurate results yielding an accuracy and F1 score of over greater than 0.70, whereas SVM with RBF kernel [222] showed the worst performance. The classifier achieved an accuracy of 0.6430 and an F1 score of 0.6714, using a gamma value of 1.0 and a C value of 1000. Also, Table A.1 reports the experimental results obtained using the HOG features. In this instance, SVM with RBF kernel achieved the highest value in terms of accuracy (0.8658) and F1 score (0.8566), considering a gamma value of 0.1 and a C value of 1000. This specific setting also outperforms the performances of the approaches that take the LBP representation. In particular, all the evaluated approaches detailed in Table A.1 achieved accuracy and F1 scores higher than 0.70.

Experiments with Convolutional Neural Networks

For the sake of comparison, we also performed object classification through the use of a Deep Convolutional Neural Network (i.e., AlexNet [100]). Considering that CNNs and Deep Learning models take advantage of a huge amount of data, we trained two CNN standard architectures (AlexNet and SmallerVGGNET), considering two different settings for the training data. First, we trained the CNNs considering the standard dataset, composed of the patches depicting the pollen objects. In a second stage, we augmented the dataset by including the segmented version of the training patches obtained by applying the segmentation mask and padding the background with all green pixels. This process can be considered an additional approach for data augmentation, which helps the CNN focus on the pollen grain in the image and ignore the remaining elements present in the background. For

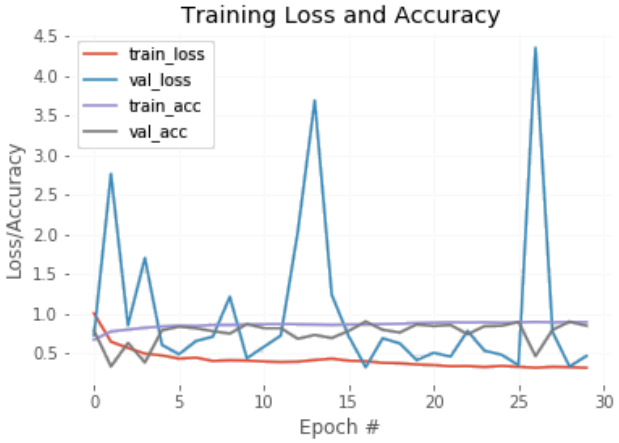


(a)

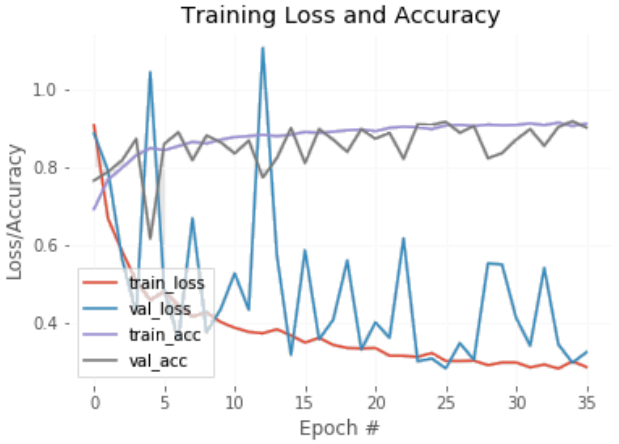
(b)

Figure A.5: AlexNet training loss and accuracy. (a) Loss/accuracy without using augmented dataset (b) Loss/accuracy with using augmented dataset.

both experiments, we also performed additional data augmentation by performing random horizontal and vertical flipping and random rotation range of 25° in training data. Moreover, we set the base learning rate to 0.001, the batch size to 64, and a number of epochs to 1000. To manage the overfitting problem, we introduced an Early stopping function, which allows us to stop training once the model performance stops improving on a hold-out validation set. After every 10 epoch, we evaluated the network performance on the test set. AlexNet yielded an average F1 score of 0.87 using the augmented dataset. With regard to the standard dataset, it could be observed that AlexNet achieved an average F1 score of 0.74, as reported in Table A.2. Fig. A.5 (a) and Fig. A.5 (b) show the performance plots related to the training of AlexNet considering the standard and the augmented dataset, respectively. In particular, each plot shows the loss and accuracy of the training and validation data over 30 epochs when Early Stopping occurred. It can be observed that the use of the augmented dataset, besides improving the model performances, helps in maintaining more stable the loss and accuracy fluctuation over the epochs



(a)



(b)

Figure A.6: SmallerVGGNET training loss and accuracy. (a) Loss/accuracy without using augmented dataset (b) Loss/accuracy with using augmented dataset.

after just 5 epochs. Table A.2 details the accuracy and F1 results measured on the test set every 10 epoch, considering the standard (SD) and the augmented dataset (AD), respectively. AlexNet outperforms the standard ML approaches only when the augmented dataset is considered. For completeness, we carry out our experiments by using a different CNN architecture, which could represent a more suitable solution for classifying images with a small size. In our case, we implemented a SmallerVGGNet, which is a variant of Very Deep Convolutional Networks (VGGNet) [161] by employing the same parameters used for AlexNet. SmallerVGGNet achieved an average F1 score of 0.85 using an augmented dataset and an average F1 score of 0.69 using images from the standard dataset. Fig. A.6 (a) and Fig. A.6 (b) show the performance plots related to the training of SmallerVGGNET considering the standard and the augmented dataset, respectively. Also, in this case, the use of the augmented dataset improves the model performances and helps in maintaining more stable the loss and accuracy fluctuation over the epochs. Table A.2 details the accuracy and F1 results measured on the test set by using the standard and the augmented dataset, respectively. SmallerVGGNET trained on the augmented dataset outperforms all the previous approaches reaching an accuracy of 0.8973 and an F1 score of 0.8914

Epoch	AlexNet - SD		SmallerVGGNET - SD	
	Accuracy	**F1_score**	**Accuracy**	**F1_score**
10	0.8067	0.7709	0.3831	0.4463
20	0.6830	0.7130	0.8428	0.8156
30	0.7992	0.7447	0.8277	0.7976
Epoch	AlexNet - AD		SmallerVGGNET - AD	
	Accuracy	**F1_score**	**Accuracy**	**F1_score**
10	0.8528	0.8507	0.8618	0.8465
20	0.8943	0.8890	**0.8973**	**0.8914**
30	**0.8963**	**0.8897**	0.8408	0.8139

Table A.2: Classification performances of AlexNet and SmallerVGGNet by using Standard Dataset (SD) and Augmented Dataset (AD).

after 20 epochs (see Table A.2). In our experiments, we observed that both AlexNet and SmallerVggnet classify pollen objects with a high confidence score. Regarding AlexNet, it tends to classify better the *Corylus Avellana* objects of category 1 (well-developed pollen) and 2 (anomalous pollen). Although AlexNet is able to classify the majority of the pollen grains accurately, it tends to misclassify pollen objects of *Corylus Avellana* (well-developed pollen) class with a texture similar to the object of *Alnus* class. Furthermore, the highest error occurred in AlexNet when the objects present features (texture) different from the peculiar ones of their class (see Fig. A.8 (a)). In this instance, AlexNet is not able to classify them accurately. Compared to AlexNet, SmallerVGGnet achieved lower accuracy and F1 score when applied to objects of categories 1 and 2. However, under challenging conditions where a pollen object is overlapped with another one (pollen or Debris), SmallerVGGNET achieved impressive results. In general, the objects of classes 2 and 4 are the most confused by the model. In Fig. A.7, examples of good classification performed by AlexNet and SmallerVGGNET are reported, respectively. In Fig. A.8 we reported examples of misclassification performed by both CNN models. Each example details the true and predicted class, as well as the confidence of the model.

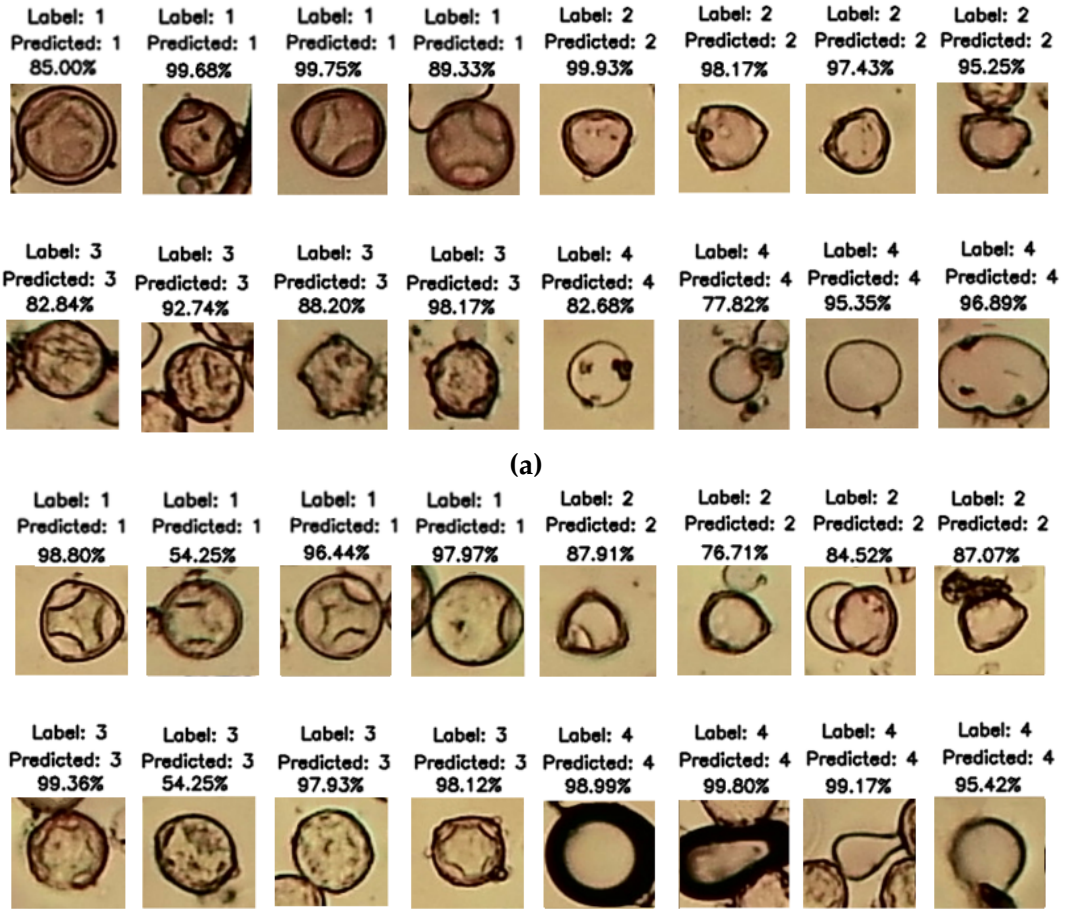


Figure A.7: Example of good classification performed by (a) AlexNet and (b) SmallerVGGNET. Each raw includes 8 examples of each involved objects correctly classified together with its confidence score.

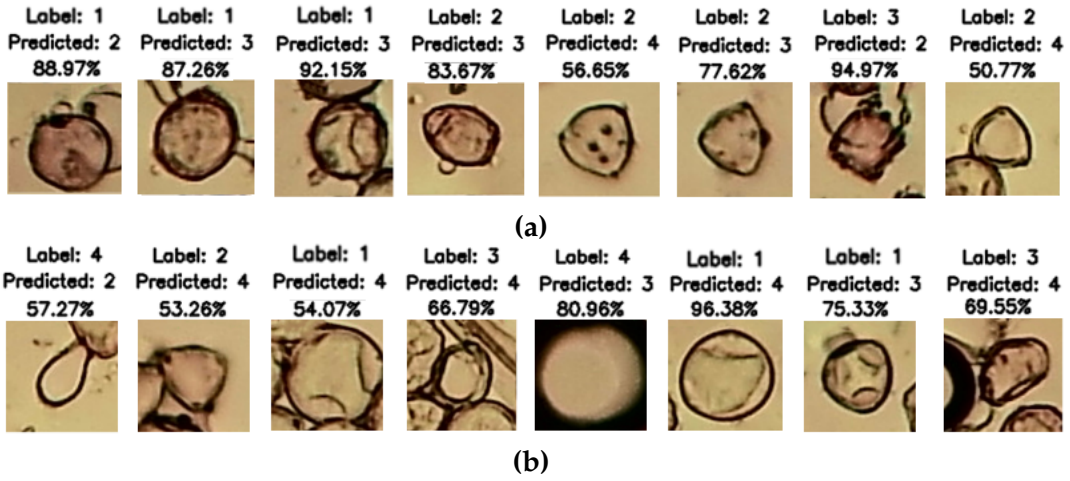


Figure A.8: Example of bad classification performed by (a) AlexNet and (b) SmallerVGGNET, together with the confidence score.

A.2 Fine-Grained Image Classification

In this section, we present an extension of our previous research activity, detailed in Section A.1. In doing so, we designed an effective pipeline to perform image classification, exploiting promising solutions which have reached state-of-art results in wide range of applications. Although our previous attempt to classify these images using DL algorithms have led to satisfactory results (see Sec. A.1), these DL approaches are still not inspire to their potential due to the challenging dataset. Motivated by these issues, we defined an innovative pipeline to improve pollen grains classification by using a Fine-Grained Visual Classification (FGVC) based approach [223]. The methods consists of a progressive training step and the application of a jigsaw patches generator in order to extract information from images at different granularity. We also implemented a Test-Time Augmentation (TTA) method to improve object classification predictions. The dataset used for the experiments is Pollen13K[3] [224], detailed in the previous section. In particular, the Pollen13K dataset includes 5 categories of objects. However, we considered the 4 classes and the train/test data splitting used during the International Pollen Grain Classification Challenge 2020. The dataset is publicly available, however, due to the nature of the competition, details about the employed methods are missing [225]. The classification of pollen objects has become a hot research topic in the field of aerobiology. Hence, the automation of pollen classification that could operate largely independently of a human operator would be of great benefit.

A.2.1 Method and Materials

In this section, we introduced the pipeline used for the classification of pollen grains. In particular, we reported more technical details regarding the progressive training strategy.

A.2.2 Training data augmentation

In [226], data augmentation is performed by operating the Cut Occlusion strategy. As mentioned previously, the strategy was shown to be effective for pollen object classification. Inspired by the results of Gui et al. [226], we reproduce the Cut Occlusion strategy in order to create new instances of the training data. The main advantage of this approach consists in avoiding some parts of the images using black patches in order to help the model extract discriminative features from the pollen wall. This strategy can bring substantial improvements for the pollen classification task, where extracting discriminative features from the aperture area of pollen grain instance seems to be more important than concentrating on the center area of the pollen object.

A.2.3 Pipeline

In Fig. A.2.3, the full pipeline is depicted. The architecture design was firstly introduced in [223], which tackles the image classification task by introducing a novel approach based on Progressive Multi-Granularity training strategy (PMG). As discussed, pollen objects from the Pollen13K dataset, belonging to different classes,

[3] more details are available on the dataset website: `https://iplab.dmi.unict.it/pollengraindataset/dataset`

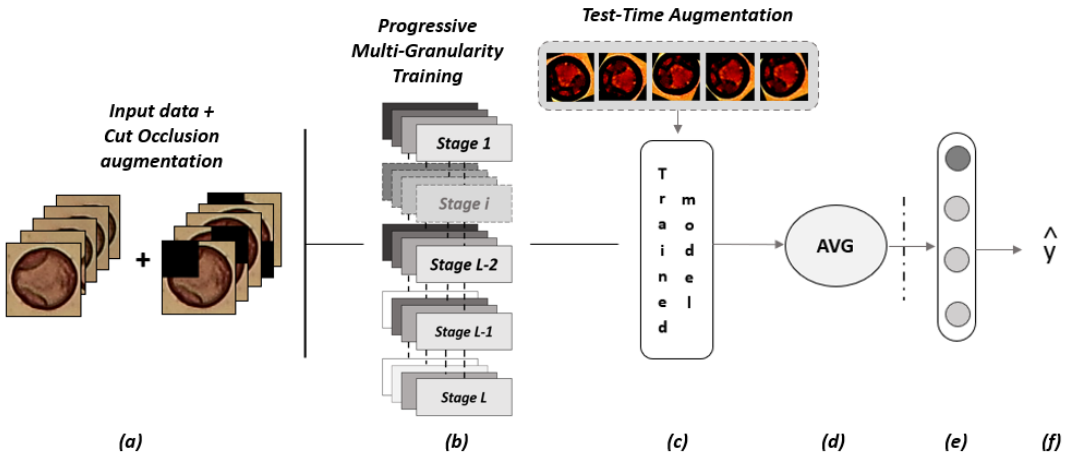


Figure A.9: The overall pipeline. (a) Input data consisting of pollen grains from Pollen13K and augmented dataset with Cut Occlusion. (b) Training performed using Progressive Multi-Granularity strategy. (c) Test-Time Augmentation. (d) Average calculation of predictions. (e) Max value of each predictions. (f) Predicted label.

present a similar appearance. In addition, objects in the same category could report varying appearances. Therefore, applying a fine-grained visual classification approach, taking advantage of local features information, could lead to remarkable improvements. The framework consists of two main components: (a) a progressive training method to add new layers during the training process to extract discriminative features from images with different granularities. Hence, the process starts at a low stage and progressively includes new layers. (b) a jigsaw patches generator [227] to capture local information from images. In [223], the authors used ResNet50 as backbone. In our study, we use ResNet101 [162] as feature extractor. For each layer $\mathcal{L}$ of the feature extractor, a new convolution layer is added, taking as input the feature maps from the output of intermediate layers that are transformed into a vector representation. Then, classification modules are added to calculate the probability distribution between classes. Finally, the outputs of the last levels are concatenated.

Progressive Training. This technique allows to train the model starting from the low stage and then to add new layers. The advantage of this technique consists in forcing the model to learn discriminative information from local details rather than focusing on global information. The loss cross-entropy function $\mathcal{L}_{CE}$ is applied to the output of each stage and the output of the concatenated features.

jigsaw puzzle generator. To train the PMG model, we define a set of jigsaw puzzle permutations. This approach has been widely employed to find the multi- granularities of the images during the training stage. Given an image x, it can be subdivided into k patches. Then, the patches are shuffled randomly and merged into a new image x'.

Test Time Augmentation. In order to boost the prediction accuracy, we implemented a strategy called Test Time Augmentation (TTA), which consists of applying data augmentation techniques to the test set in order to improve the prediction of a given class of objects. For this reason, we create several variants of the same image, applying a horizontal or vertical flip, standard color augmentation, or other geometric transformations. We computed a prediction for each of these images. Then, we

average these predictions and calculate the max value in order to obtain which prediction has the highest confidence score. Finally, we computed the predicted class for the analyzed object. By applying this strategy, we avoid the uncertain of the model by averaging the predictions and averaging the error. In this context, we created 5 different variants for every single image of the test set by applying a horizontal flip, a vertical flip, and a random rotation. The rotation angle ranges from -90° to 90°.In addition, as in the training set, we performed an image resize of 550×550 pixels and a center crop of 448×448 pixels. Moreover, we normalized the data set with a mean and a standard deviation of 0.5.

A.2.4 Experiments

To evaluate the performance of the proposed pipeline, we performed rigorous experiments on two image datasets: Pollen13K [224] and Augmented Pollen13K. We implemented the proposed approaches as well as several state-of-the-art approaches on these datasets. The experimental settings are given in the following subsections.

A.2.5 Dataset

Original dataset. We used the dataset Pollen13K [224] which has been introduced on the previous section (Sec. A.1).
Augmented Dataset. In order to boost the predictions, we performed dataset augmentation by using Cut Occlusion strategy [226]. We inserted occlusions around the center of pollen area. In particular, we created 4 different variants for each image of the training set.

Implementation Details of the proposed pipeline

All experiments were conducted using PyTorch [163] over a cluster of GPU NVIDIA T4. We employed ResNet101 [162] as backbone. All setting are indicated in [223], where $S = 3$, $\alpha = 1$, and $\beta = 2$. In addition, the input images were resized to 550×550 pixels and randomly cropped by 448×448 pixels. A random horizontal flipping is applied for data augmentation for training data. We use Stochastic Gradient Descent (SGD) [164] optimizer and batch normalization as the regularizer. We train the model for 100 epochs. The batch size was set to 16. Moreover, we used a weight decay of 0.0005 and a momentum of 0.9.

A.2.6 Other experiments

Dealing with imbalanced data. In this research activity, we evaluate the performance of the aforementioned algorithms using the Pollen13K dataset [224]. To the best of our knowledge, it represents the public dataset with the largest number of pollen objects, with more than $13,000$ objects. However, the Pollen13K dataset consists of imbalanced classes since the largest class consists of the objects from class Alnus ($8,216$ objects in the train set). Other classes include a total number of objects less than $1,600$. Motivated by these issues, we provided an effective solution by implementing the Weighted Random Sampler (WRS) function to deal with imbalanced datasets and preventing overfitting problems. Hence, one of the proposed solutions is to oversample minority classes [228]. By applying this technique, we balanced batches of data. As a result, during training time, the model will not concentrate significantly on one class over another, and risks of overfitting are reduced. Basically, the WRS method uses the array of weights that corresponds to weights given

Method	Accuracy	F1 (weighted)	F1 (macro)
WRS + Resnet101	92.667 %	92.899 %	88.233 %
WRS + ResAttNet	83.776 %	84.497 %	77.627 %
WRS + CutMix + ResNet101	93.922 %	94.046 %	90.097 %
PMG [223]	**96.384 %**	**96.349 %**	**93.585 %**
Method + TTA	Accuracy	F1 (weighted)	F1 (macro)
WRS + Resnet101 + TTA	94.626 %	94.702 %	91.266 %
WRS + ResAttNet + TTA	83.626 %	84.748 %	80.416 %
WRS + CutMix + ResNet101 + TTA	95.228 %	95.280 %	92.581 %
CutOcclusion + PMG + TTA (Proposed)	**97.087 %**	**97.050 %**	**94.726 %**

Table A.3: Comparison between DL approaches. On the top part of the table, we reported evaluation results without applying TTA. On the bottom part of the table, we reported results by applying TTA method.

to each class. The goal is to assign a higher weight to the minor class, providing a more robust classification. Finally, we evaluate the performance of each classifier by also using the weighted and macro F1 score, which represent two more reliable performance metrics than accuracy. The weighted F1 score function calculates the F1 metrics for each class and their average weighted by support (i.e., the number of true instances for each class). The F1 macro score computes the F1 for each label and returns the average.

Implementation Details. We use the Pytorch [163] Deep Learning library for performing the experiments detailed below. We resized input images by 256×256 pixels. A 224×224 center crop is sampled from an augment image, applying geometric transformations. The network is trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9. We set the initial learning rate to 0.0001, decaying the learning rate by a factor of 0.1 every 7 epochs. We set the number of epochs to 100. All the experiments use a batch size of 16. With regard to the methods based on CutMix strategy [229], we set hyperparameters values to $\beta = 1.0$ and *cutmix probability* to 0.5.

A.2.7 Results and Discussion

This section presents the results obtained for a pool of DL algorithms, providing a benchmarking evaluation of the performance of the techniques herein proposed for pollen grains classification. In Table A.3, results show that PMG [223] method lead to a boost of the prediction accuracy than other methods. The main advantage of this approach includes the analysis of images with different granularities. Basically, it forces each stage of the network to focus on local features rather than on global information. Furthermore, a jigsaw generator performs an image splitting into several patches during the training phase, providing discriminative information at the specific granularity level. Although the method based on Residual Attention Network (ResAttNet) produces remarkable results, it fails to outperform the other proposed methods. In both cases, we observed that the CutMix-based approach leads to better classification results than the pre-trained model (ResNet101) without using this strategy as data augmentation. According to these results, the method that yields good results, both in terms of accuracy and F1 scores, is the approach based on progressive training and a jigsaw generator, i.e., PMG. To further improve the performance of the PMG model, we defined a data augmentation technique based

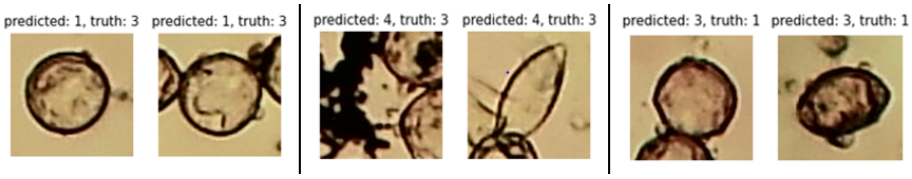


Figure A.10: Example of bad classification performed by PMG. These objects are classified accurately by PMG with training augmentation and TTA.

on Cut Occlusion and a Test Time Augmentation to achieve higher accuracy during inference. The proposed pipeline yield better results than other methods, confirming the effectiveness of the proposed framework. This strategy tends to improve the classification results and obtain results consistent with state-of-the-art. We reported the achieved results in Table A.3. As observed, the TTA strategy provides reliable results in terms of accuracy and F1-score (weighted and macro) than other methods where this strategy was not applied to. With regard to Residual Attention Network (ResAttNet), the accuracy value is decreased compared to the value from the previous experiment. Instead, the F1-score metrics improve their value. In general, we observe that DL methods combined with the TTA strategy yields better results than methods not using Test Time Augmentation strategy.

Misclassification: In Fig. A.2.7, we report some examples of misclassification obtained by the standard PMG [223] algorithm without using training augmentation and TTA approach. As observed, the standard PMG approach [223] misclassifies objects of class 1 with objects of class 3 and vice versa. In fact, the objects from these classes present similar characteristics, leading to a challenging image classification task. Furthermore, objects belonging to class 4 are indicated as class 3 objects. Probably, it depends on the presence of other objects within the image, which fools the model, forcing it to extract their features and leading to misclassification. However, the implementation of the Cut Occlusion strategy allows us to avoid these objects, encouraging the network to focus on the object depicted at the center of the image, reporting better classification predictions.

Comparisons with previous studies. We also reported a comparison between our proposed approach and previous studies for the classification of pollen grains. Fang et al. [230] propose a blending strategy consisting of a Destruction and Construction Learning architecture [231] and DenseNAS [232] output vectors to be used as the input of a Random Forest Classifier, which performs the final classification. Gui et al. [226] generated several images by applying the Cut Occlusion approach. The trained model is based on ResNet101. In our previous study (see Sec. A.1, we investigated the performance of several Machine Learning approaches, such as AlexNet, SmallerVGGNet, etc. Fang et al. [230] leads to the best results in terms of accuracy (97.539) and F1-score (97.510), whereas Penghui Gui et al. achieved an accuracy of 97.290 and an F1-score of 97.260. Our previous method [233] achieved an accuracy of 89.730 % and an F1-score of 89.140%. In our experiments by using the proposed pipeline, we achieved an accuracy of 97.087 % and an F1-score (weighted) of 97.050 %, providing results similar to [230] and [226]. We also performed cross-validation, achieving good results in accuracy (96.5%) and F1-score (96%). Although the experiments suggest that the algorithm we proposed provides the best results, we did not perform an accurate validation step. On the contrary, our 3-fold cross-validation method has proved to reach more robust results.

Appendix B

Deep Learning for Quantitative Finance

This section presents additional works done by the Ph.D. candidate. The described research activity encompasses different approaches and perspectives related to the Quantitative Finance field. Since such works are not relevant for the topic addressed in this dissertation, we only reported the abstracts of the proposed research activities.

B.1 Overview

The analysis of financial data represents a challenge that researchers had to deal with. The rethinking of the basis of financial markets has led to an urgent demand for developing innovative models to understand financial assets. In the past few years, researchers have proposed several systems based on traditional approaches, such as auto-regressive integrated moving average (ARIMA) and the exponential smoothing model, in order to devise an accurate data representation. In this regard, researchers have modeled a convenient representation for financial data, the so-called time-series (i.e., numerical data points observed sequentially through time). However, the scientific community has highlighted the difficulty of studying financial time series accurately due to their non-linear and non-stationary patterns. Motivated by these issues, we investigated the problem of processing financial data by using Dl techniques. In recent years there has been increasing interest in predicting the future behavior of complex systems by involving a temporal component [234]. Despite skepticism about the effectiveness of these approaches, researchers have proven that DL methods outperform traditional approaches by effectively identify significant data information from irrelevant ones. In this section, it is reported a brief introduction for each of our research activities related to the Quantitative Finance field.

B.1.1 Trading System: a Markov-based Machine Learning framework

Stock market prediction and trading have attracted the effort of many researchers in several scientific areas because it is a challenging task due to the high complexity of the market. More investors put their effort into developing a systematic approach, i.e., the so-called "Trading System (TS)" for stock pricing and trend prediction. The introduction of the Trading On-Line (TOL) has significantly improved the overall number of daily transactions on the stock market with the consequent increase of the market complexity and liquidity. One of the main consequences of the TOL is the "automatic trading," i.e., an ad-hoc algorithmic robot able to automatically analyze a lot of financial data with a target to open/close several trading operations in

such reduced time for increasing the profitability of the trading system. When the number of such automatic operations increases significantly, the trading approach is known as High-Frequency Trading (HFT). In this context, the usage of ML methods has improved the robustness of the trading systems, including the HFT sector. We propose an innovative approach based on the usage of the ad-hoc ML approach, starting from historical data analysis, which is able to perform careful stock price prediction. The stock price prediction accuracy is further improved by using adaptive correction based on the hypothesis that stock price formation is regulated by Markov stochastic propriety. The validation results applied to such shares and financial instruments confirm the robustness and effectiveness of the proposed automatic trading algorithm. More details are reported on [79].

B.1.2 Grid trading system robot (gtsbot)

Grid algorithmic trading has become quite popular among traders because it shows several advantages with respect to similar approaches. Basically, a grid trading strategy is a method that seeks to make a profit on the market movements of the underlying financial instrument by positioning buy and sell orders properly time-spaced (grid distance). The main advantage of the grid trading strategy is the financial sustainability of the algorithm because it provides a robust way to mediate losses in financial transactions, even though this also means a very complicated trades management algorithm. For these reasons, grid trading is certainly one of the best approaches to be used in high-frequency trading (HFT) strategies. Due to the high level of unpredictability of the financial markets, many investment funds and institutional traders are opting for the HFT (high-frequency trading) systems, which allow them to obtain high performance due to the large number of financial transactions executed in the short-term timeframe. The combination of HFT strategies with the use of ML methods for the financial time series forecast has significantly improved the capability and overall performance of modern automated trading systems. We propose an automatic HFT grid trading system that operates in the FOREX (foreign exchange) market. The performance of the proposed algorithm, together with the reduced drawdown, confirmed the effectiveness and robustness of the proposed approach. More details are reported on [86].

B.1.3 Machine learning for quantitative finance applications: A survey

In the proposed survey, we selected studies and research works based on the ML approach or classical method in order to analyze time series in the financial domain. We shed light on the promising results achieved by ML approaches for time-series forecasting problems. Unlike relevant existing review articles [235], the survey not only focuses on summarizing several approaches suitable for solving financial market problems but also compares ML methods and traditional ones in order to discuss which method could be more effective considering the challenging financial scenario. Also, we provided results of selected studies to highlight the better overall performance of ML-based systems over traditional ones. More details are reported on [236].

Bibliography

[1] M. Bahri, A. Bifet, J. Gama, H. M. Gomes, and S. Maniu, "Data stream analysis: Foundations, major tasks and tools," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 3, e1405, 2021.

[2] M. Winikoff, "Towards trusting autonomous systems," in *International Workshop on Engineering Multi-Agent Systems*, Springer, 2017, pp. 3–20.

[3] A. Bener, E. Yildirim, T. Özkan, and T. Lajunen, "Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population based case and control study," *Journal of Traffic and Transportation engineering (English edition)*, vol. 4, no. 5, pp. 496–502, 2017.

[4] T. Aggarwal, A. Furqan, and K. Kalra, "Feature extraction and lda based classification of lung nodules in chest ct scan images," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2015, pp. 1189–1193.

[5] H. Shaziya, K. Shyamala, and R. Zaheer, "Automatic lung segmentation on thoracic ct scans using u-net convolutional network," in *2018 International conference on communication and signal processing (ICCSP)*, IEEE, 2018, pp. 0643–0647.

[6] *Road accidents report*, `https://www.nsc.org/road-safety/safety-topics/fatality-estimates`, Accessed: 2020-08-06.

[7] F. Trenta, S. Conoci, F. Rundo, and S. Battiato, "Advanced motion-tracking system with multi-layers deep learning framework for innovative car-driver drowsiness monitoring," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, IEEE, 2019, pp. 1–5.

[8] F. Rundo, S. Conoci, A. Ortis, and S. Battiato, "An advanced bio-inspired photoplethysmography (ppg) and ecg pattern recognition system for medical assessment," *Sensors*, vol. 18, no. 2, p. 405, 2018.

[9] F. Rundo, S. Rinella, S. Massimino, M. Coco, G. Fallica, R. Parenti, S. Conoci, and V. Perciavalle, "An innovative deep learning algorithm for drowsiness detection from eeg signal," *Computation*, vol. 7, no. 1, p. 13, 2019.

[10] V. Vinciguerra, E. Ambra, L. Maddiona, M. Romeo, M. Mazzillo, F. Rundo, G. Fallica, F. di Pompeo, A. M. Chiarelli, F. Zappasodi, *et al.*, "Ppg/ecg multisite combo system based on sipm technology," in *Convegno Nazionale Sensori*, Springer, 2018, pp. 353–360.

[11] M. Mazzillo, L. Maddiona, F. Rundo, A. Sciuto, S. Libertino, S. Lombardo, and G. Fallica, "Characterization of sipms with nir long-pass interferential and plastic filters," *IEEE Photonics Journal*, vol. 10, no. 3, pp. 1–12, 2018.

[12] G. Grasso, P. Perconti, and A. Plebe, "Assessing social driving behavior," in *International Conference on Intelligent Human Systems Integration*, Springer, 2019, pp. 111–115.

[13] G. M. Grasso, C. Lucifora, P. Perconti, and A. Plebe, "Evaluating mentalization during driving.," in *VEHITS*, 2019, pp. 536–541.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*, IEEE, 2016, pp. 565–571.

[15] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06-083*, 2017.

[17] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *arXiv preprint*, arXiv:1805.12152, 2018.

[18] M. Chui, J. Manyika, and M. Miremadi, "Where machines could replace humans and where they can't (yet)," 2016.

[19] P. Philip, "Sleepiness of occupational drivers," *Industrial health*, vol. 43, no. 1, pp. 30–33, 2005.

[20] W. Vanlaar, H. Simpson, D. Mayhew, and R. Robertson, "Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors," *Journal of safety research*, vol. 39, no. 3, pp. 303–309, 2008.

[21] T. Igasaki, K. Nagasawa, N. Murayama, and Z. Hu, "Drowsiness estimation under driving environment by heart rate variability and/or breathing rate variability with logistic regression analysis," in *2015 8th International Conference on Biomedical Engineering and Informatics (BMEI)*, IEEE, 2015, pp. 189–193.

[22] B. Abi-Saleh and B. Omar, "Einthoven's triangle transparency: A practical method to explain limb lead configuration following single lead misplacements," *Reviews in cardiovascular medicine*, vol. 11, no. 1, pp. 33–38, 2019.

[23] F. Rundo, S. Petralia, G. Fallica, and S. Conoci, "A nonlinear pattern recognition pipeline for ppg/ecg medical assessments," in *Convegno Nazionale Sensori*, Springer, 2018, pp. 473–480.

[24] F. Rundo, C. Spampinato, and S. Conoci, "Ad-hoc shallow neural network to learn hyper filtered photoplethysmographic (ppg) signal for efficient car-driver drowsiness monitoring," *Electronics*, vol. 8, no. 8, p. 890, 2019.

[25] S. Conoci, F. Rundo, G. Fallica, D. Lena, I. Buraioli, and D. Demarchi, "Live demonstration of portable systems based on silicon sensors for the monitoring of physiological parameters of driver drowsiness and pulse wave velocity," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2018, pp. 1–3.

[26] K. Ogawa and M. Shimotani, "A drowsiness detection system," *Mitsubishi Electric Advance*, pp. 13–16, 1997.

[27] L. M. Bergasa, R. Barea, E. L. Guillén, M. S. Escudero, L. Boquete, and J. I. Pinedo, "Facial features tracking applied to drivers drowsiness detection.," in *Applied Informatics*, 2003, pp. 231–236.

[28] I. Park, J.-H. Ahn, and H. Byun, "Efficient measurement of eye blinking under various illumination conditions for drowsiness detection systems," in *18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, vol. 1, 2006, pp. 383–386.

[29] T. Wang and P. Shi, "Yawning detection for determining driver drowsiness," in *Proceedings of 2005 IEEE International Workshop on VLSI Design and Video Technology, 2005.*, IEEE, 2005, pp. 373–376.

[30] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," in *International Workshop on Human-Computer Interaction*, Springer, 2007, pp. 6–18.

[31] J. Batista, "A drowsiness and point of attention monitoring system for driver vigilance," in *2007 IEEE Intelligent Transportation Systems Conference*, IEEE, 2007, pp. 702–708.

[32] F. Friedrichs and B. Yang, "Camera-based drowsiness reference for driver state classification under real driving conditions," in *2010 IEEE Intelligent Vehicles Symposium*, IEEE, 2010, pp. 101–106.

[33] A. A. Lenskiy and J.-S. Lee, "Driver's eye blinking detection using novel color and texture segmentation algorithms," *International journal of control, automation and systems*, vol. 10, no. 2, pp. 317–327, 2012.

[34] T. Hong and H. Qin, "Drivers drowsiness detection in embedded system," in *2007 IEEE International Conference on Vehicular Electronics and Safety*, IEEE, 2007, pp. 1–5.

[35] G. Zhenhai, L. DinhDat, H. Hongyu, Y. Ziwen, and W. Xinyu, "Driver drowsiness detection based on time series analysis of steering wheel angular velocity," in *2017 9th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, IEEE, 2017, pp. 99–101.

[36] Z. Li, S. E. Li, R. Li, B. Cheng, and J. Shi, "Online detection of driver fatigue using steering wheel angles for real driving conditions," *Sensors*, vol. 17, no. 3, p. 495, 2017.

[37] A. D. McDonald, C. Schwarz, J. D. Lee, and T. L. Brown, "Real-time detection of drowsiness related lane departures using steering wheel angle," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 56, 2012, pp. 2201–2205.

[38] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, "Detection of driver's drowsiness by means of hrv analysis," in *2011 Computing in Cardiology*, IEEE, 2011, pp. 89–92.

[39] M. Szypulska and Z. Piotrowski, "Prediction of fatigue and sleep onset using hrv analysis," in *Proceedings of the 19th International Conference Mixed Design of Integrated Circuits and Systems-MIXDES 2012*, IEEE, 2012, pp. 543–546.

[40] H. Lee, J. Lee, and M. Shin, "Using wearable ecg/ppg sensors for driver drowsiness detection based on distinguishable pattern of recurrence plots," *Electronics*, vol. 8, no. 2, p. 192, 2019.

[41] G.-S. Ryu, J. You, V. Kostianovskii, E.-B. Lee, Y. Kim, C. Park, and Y.-Y. Noh, "Flexible and printed ppg sensors for estimation of drowsiness," *IEEE Transactions on Electron Devices*, vol. 65, no. 7, pp. 2997–3004, 2018.

[42] D. Kurian, J. J. PL, K. Radhakrishnan, and A. A. Balakrishnan, "Drowsiness detection using photoplethysmography signal," in *2014 Fourth international conference on advances in computing and communications*, IEEE, 2014, pp. 73–76.

[43] G. Li and W.-Y. Chung, "Detection of driver drowsiness using wavelet analysis of heart rate variability and a support vector machine classifier," *Sensors*, vol. 13, no. 12, pp. 16 494–16 511, 2013.

[44] H. Park, S. Oh, and M. Hahn, "Drowsy driving detection based on human pulse wave by photoplethysmography signal processing," in *Proceedings of the 3rd International Universal Communication Symposium*, 2009, pp. 89–92.

[45] N. N. Sari and Y.-P. Huang, "A two-stage intelligent model to extract features from ppg for drowsiness detection," in *2016 International Conference on System Science and Engineering (ICSSE)*, IEEE, 2016, pp. 1–2.

[46] B.-G. Lee, S.-J. Jung, and W.-Y. Chung, "Real-time physiological and vision monitoring of vehicle driver for non-intrusive drowsiness detection," *IET communications*, vol. 5, no. 17, pp. 2461–2469, 2011.

[47] B.-G. Lee and W.-Y. Chung, "Multi-classifier for highly reliable driver drowsiness detection in android platform," *Biomedical Engineering: Applications, Basis and Communications*, vol. 24, no. 02, pp. 147–154, 2012.

[48] S.-P. Cheon and S.-J. Kang, "Sensor-based driver condition recognition using support vector machine for the detection of driver drowsiness," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2017, pp. 1517–1522.

[49] H.-T. Choi, M.-K. Back, and K.-C. Lee, "Driver drowsiness detection based on multimodal using fusion of visual-feature and bio-signal," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2018, pp. 1249–1251.

[50] S. I. Tonni, T. A. Aka, M. M. Antik, K. A. Taher, M. Mahmud, and M. S. Kaiser, "Artificial intelligence based driver vigilance system for accident prevention," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, IEEE, 2021, pp. 412–416.

[51] E. Monte-Moreno, "Non-invasive estimate of blood glucose and blood pressure from a photoplethysmograph by means of machine learning techniques," *Artificial intelligence in medicine*, vol. 53, no. 2, pp. 127–138, 2011.

[52] G. Slapničar, N. Mlakar, and M. Luštrek, "Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network," *Sensors*, vol. 19, no. 15, p. 3420, 2019.

[53] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired?" *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008.

[54] S. Romdhani, P. Torr, B. Scholkopf, and A. Blake, "Computationally efficient face detection," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, IEEE, vol. 2, 2001, pp. 695–700.

[55] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, IEEE, vol. 1, 2001, pp. I–I.

[56] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *International Journal of Neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.

[57] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, Springer, 2006, pp. 404–417.

[58] A. Siennicka, D. Quintana, P. Fedurek, A. Wijata, B. Paleczny, B. Ponikowska, and D. Danel, "Resting heart rate variability, attention and attention maintenance in young adults," *International Journal of Psychophysiology*, vol. 143, pp. 126–131, 2019.

[59] A. E. Dastjerdi, M. Kachuee, and M. Shabany, "Non-invasive blood pressure estimation using phonocardiogram," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2017, pp. 1–4.

[60] B. S. Dangra, D. Rajput, M. Bedekar, and S. S. Panicker, "Profiling of automobile drivers using car games," in *2015 International Conference on Pervasive Computing (ICPC)*, IEEE, 2015, pp. 1–5.

[61] K. Kim, H. Choi, and B. Jang, "Design of the Driver-Adaptive Vehicle Interaction System," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, IEEE, 2018, pp. 297–299.

[62] M. M. Mubasher, S. W. Jaffry, and R. Jahangir, "Modeling of individual differences in car-following behaviour of drivers," in *2017 International Multi-topic Conference (INMIC)*, IEEE, 2017, pp. 1–7.

[63] G. Castignani and R. Frank, "SenseFleet: A smartphone-based driver profiling platform," in *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, IEEE, 2014, pp. 144–145.

[64] F. Rundo, "Deep lstm with dynamic time warping processing framework: A novel advanced algorithm with biosensor system for an efficient car-driver recognition," *Electronics*, vol. 9, no. 4, p. 616, 2020.

[65] M. Altun and M. Celenk, "Road scene content analysis for driver assistance and autonomous driving," *IEEE transactions on intelligent transportation systems*, vol. 18, no. 12, pp. 3398–3407, 2017.

[66] T. Deng, K. Yang, Y. Li, and H. Yan, "Where does the driver look? top-down-based saliency detection in a traffic driving environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2051–2062, 2016.

[67] J. Ferreira, E. Carvalho, B. V. Ferreira, C. de Souza, Y. Suhara, A. Pentland, and G. Pessin, "Driver behavior profiling: An investigation with different smartphone sensors and machine learning," *PLoS one*, vol. 12, no. 4, e0174959, 2017.

[68] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.

[69] F. Rundo, F. Trenta, S. Conoci, and S. Battiato, *Image processing method and corresponding system*, US Patent App. 16/729,879, 2020.

[70] K. Fujiwara, E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda, *et al.*, "Heart rate variability-based driver drowsiness detection and its validation with eeg," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1769–1778, 2018.

[71] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[72] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[73] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[74] H.-S. Shin, S.-J. Jung, J.-J. Kim, and W.-Y. Chung, "Real time car driver's condition monitoring system," in *SENSORS, 2010 IEEE*, IEEE, 2010, pp. 951–954.

[75] Y.-P. Huang, N. N. Sari, and T.-T. Lee, "Early detection of driver drowsiness by wpt and flfnn models," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2016, pp. 000 463–000 468.

[76] F. Rundo, C. Spampinato, S. Conoci, F. Trenta, and S. Battiato, "Deep biosensing embedded system for a robust car-driving safety assessment," in *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, IEEE, 2020, pp. 1–6.

[77] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[78] G. L. B. F. Rundo S. Conoci, "Image processing method, corresponding system and computer program product," IT102018000010833, 2018.

[79] F. Rundo, F. Trenta, A. L. Di Stallo, and S. Battiato, "Advanced markov-based machine learning framework for making adaptive trading system," *Computation*, vol. 7, no. 1, p. 4, 2019.

[80] F. Rundo, C. Spampinato, S. Battiato, F. Trenta, and S. Conoci, "Advanced 1d temporal deep dilated convolutional embedded perceptual system for fast car-driver drowsiness monitoring," in *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, IEEE, 2020, pp. 1–6.

[81] F. Rundo, F. Trenta, R. Leotta, C. Spampinato, V. Piuri, S. Conoci, R. D. Labati, F. Scotti, and S. Battiato, "Advanced temporal dilated convolutional neural network for a robust car driver identification.," in *ICPR Workshops (8)*, 2020, pp. 184–199.

[82] C.-I. Chang, *Hyperspectral imaging: techniques for spectral detection and classification*. Springer Science & Business Media, 2003, vol. 1.

[83] G. Bianchi and R. Sorrentino, *Electronic filter simulation & design*. McGraw Hill Professional, 2007.

[84] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[85] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[86] F. Rundo, F. Trenta, A. L. di Stallo, and S. Battiato, "Grid trading system robot (gtsbot): A novel mathematical algorithm for trading fx market," *Applied Sciences*, vol. 9, no. 9, p. 1796, 2019.

[87] F. Rundo, S. Conoci, G. L. Banna, A. Ortis, F. Stanco, and S. Battiato, "Evaluation of levenberg–marquardt neural networks and stacked autoencoders clustering for skin lesion analysis, screening and follow-up," *IET Computer Vision*, vol. 12, no. 7, pp. 957–962, 2018.

[88] *Stmicroelectronics accordo 5*, `https://www.st.com/en/automotive-infotainment-and-telematics/automotive-infotainment-socs.html?icmp=tt4379_gl_pron_nov2016`, Accessed: 2020-08-06.

[89] *Stmicroelectronics mcus*, https://www.st.com/en/automotive-infotainment-and-telematics/automotive-infotain-ment-socs.html?icmp=tt4379_gl_pron_nov2016, Accessed: 2020-08-06.

[90] F. Rundo, S. Conoci, S. Battiato, F. Trenta, and C. Spampinato, "Innovative saliency based deep driving scene understanding system for automatic safety assessment in next-generation cars," in *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, IEEE, 2020, pp. 1–6.

[91] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[92] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2394–2403.

[93] S. Battiato, S. Conoci, R. Leotta, A. Ortis, F. Rundo, and F. Trenta, "Benchmarking of computer vision algorithms for driver monitoring on automotive-grade devices," in *2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, IEEE, 2020, pp. 1–6.

[94] A. Torrisi, G. M. Farinella, G. Puglisi, and S. Battiato, "Selecting discriminative clbp patterns for age estimation," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2015, pp. 1–6.

[95] L. K. McIntire, R. A. McKinley, C. Goodyear, and J. P. McIntire, "Detection of vigilance performance using eye blinks," *Applied ergonomics*, vol. 45, no. 2, pp. 354–362, 2014.

[96] J. Cech and T. Soukupova, "Real-time eye blink detection using facial landmarks," *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pp. 1–8, 2016.

[97] P. Viola, M. Jones, *et al.*, "Robust real-time object detection," *International journal of computer vision*, vol. 4, no. 34-47, p. 4, 2001.

[98] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[99] *Yocto overview manual*, https://www.yoctoproject.org/docs/3.1.2/overview-manual/overview-manual.html, Accessed: 2020-05-15.

[100] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[101] H. Becker, W. Nettleton, P. Meyers, J. Sweeney, and C. Nice, "Digital computer determination of a medical diagnostic index directly from chest x-ray images," *IEEE Transactions on Biomedical Engineering*, no. 3, pp. 67–72, 1964.

[102] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer.," *Radiology*, vol. 187, no. 1, pp. 81–87, 1993.

[103] S.-C. Lo, S.-L. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE transactions on medical imaging*, vol. 14, no. 4, pp. 711–718, 1995.

[104] M. T. Vlaardingerbroek and J. A. Boer, *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media, 2013.

[105] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[106] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

[107] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, *et al.*, "Interpretability of deep learning models: A survey of results," in *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, IEEE, 2017, pp. 1–6.

[108] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[109] F. Wang, R. Kaushal, and D. Khullar, *Should health care demand interpretable artificial intelligence or accept "black box" medicine?* 2020.

[110] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[111] M. T. Ribeiro, S. Singh, and C. Guestrin, ""' why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[112] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[113] M. Reyes, R. Meier, S. Pereira, C. A. Silva, F.-M. Dahlweid, H. v. Tengg-Kobligk, R. M. Summers, and R. Wiest, "On the interpretability of artificial intelligence in radiology: Challenges and opportunities," *Radiology: Artificial Intelligence*, vol. 2, no. 3, e190043, 2020.

[114] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[115] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[116] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[117] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[118] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[119] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[120] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[121] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques," *Computational Intelligence*, 2020.

[122] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," *arXiv preprint arXiv:1703.03717*, 2017.

[123] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*, PMLR, 2017, pp. 18-85–1894.

[124] C. Anil, J. Lucas, and R. Grosse, "Sorting out lipschitz function approximation," in *International Conference on Machine Learning*, PMLR, 2019, pp. 291–301.

[125] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Thirty-second AAAI Conference*, 2018.

[126] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[127] A. Chan, Y. Tay, Y. S. Ong, and J. Fu, "Jacobian adversarially regularized networks for robustness," *arXiv preprint arXiv:1912.10185*, 2019.

[128] D. Jakubovitz and R. Giryes, "Improving dnn robustness to adversarial attacks using jacobian regularization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 514–529.

[129] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2147–2154.

[130] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[131] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *ICLR*, 2019.

[132] C. Etmann, S. Lunz, P. Maass, and C. Schoenlieb, "On the connection between adversarial robustness and saliency map interpretability," in *ICML 2019*.

[133] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on mnist," *arXiv preprint arXiv:1805.09190*, 2018.

[134] P. Maini, E. Wong, and Z. Kolter, "Adversarial robustness against the union of multiple perturbation models," in *International Conference on Machine Learning*, PMLR, 2020, pp. 6640–6650.

[135] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," *arXiv preprint arXiv:2010.00467*, 2020.

[136] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," *arXiv preprint arXiv:2102.01356*, 2021.

[137] B. Kim, J. Seo, and T. Jeon, "Bridging adversarial robustness and gradient interpretability," *arXiv preprint arXiv:1903.11626*, 2019.

[138] L. J. Ba and R. Caruana, "Do deep nets really need to be deep?" *arXiv preprint arXiv:1312.6184*, 2013.

[139] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[140] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020.

[141] C. Pino, S. Palazzo, F. Trenta, F. Cordero, U. Bagci, F. Rundo, S. Battiato, D. Giordano, M. Aldinucci, and C. Spampinato, "Interpretable deep model for predicting gene-addicted non-small-cell lung cancer in ct scans," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2021, pp. 891–894.

[142] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *CVPRW 2017*, IEEE, 2017, pp. 1175–1183.

[143] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks.," in *CVPR*, vol. 1, 2017, p. 3.

[144] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[145] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[146] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 149–157.

[147] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv:1312.6199*, 2013.

[148] J. Hoffman, D. A. Roberts, and S. Yaida, "Robust learning with jacobian regularization," *arXiv preprint arXiv:1908.02729*, 2019.

[149] B. Simpson, F. Dutil, Y. Bengio, and J. P. Cohen, "Gradmask: Reduce overfitting by regularizing saliency," *arXiv preprint arXiv:1904.07478*, 2019.

[150] S. Singh, K. Ho-Shon, S. Karimi, and L. Hamey, "Modality classification and concept detection in medical images using deep transfer learning," in *2018 International conference on image and vision computing New Zealand (IVCNZ)*, IEEE, 2018, pp. 1–9.

[151] M. Hassan, S. Ali, H. Alquhayz, and K. Safdar, "Developing intelligent medical image modality classification system using deep transfer learning and lda," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[152] I. Kitanovski, K. Trojacanec, I. Dimitrovski, and S. Loskovska, "Modality classification using texture features," in *International Conference on ICT Innovations*, Springer, 2011, pp. 189–198.

[153] J. Kalpathy-Cramer, W. Hersh, *et al.*, "Automatic image modality based classification and annotation to improve medical image retrieval," in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, IOS Press, 2007, p. 1334.

[154] M. Y. Khachane and R. Ramteke, "Modality based medical image classification," in *Emerging research in computing, information, communication and applications*, Springer, 2016, pp. 597–606.

[155] X.-H. Han and Y.-W. Chen, "Biomedical imaging modality classification using combined visual features and textual terms," *International journal of biomedical imaging*, vol. 2011, 2011.

[156] C.-H. Chiang, C.-L. Weng, and H.-W. Chiu, "Automatic classification of medical image modality and anatomical location using convolutional neural network," *Plos one*, vol. 16, no. 6, e0253205, 2021.

[157] G. van Tulder and M. de Bruijne, "Learning cross-modality representations from multi-modal images," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 638–648, 2018.

[158] D. Cheng and M. Liu, "Cnns based multi-modality classification for ad diagnosis," in *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, IEEE, 2017, pp. 1–5.

[159] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, *et al.*, "Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease," *IEEE transactions on biomedical engineering*, vol. 62, no. 4, pp. 1132–1140, 2014.

[160] J. Arias, J. Martinez-Gomez, J. A. Gamez, A. G. S. de Herrera, and H. Müller, "Medical image modality classification using discrete bayesian networks," *Computer vision and image understanding*, vol. 151, pp. 61–71, 2016.

[161] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[163] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[164] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

[165] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," *arXiv preprint arXiv:2102.06171*, 2021.

[166] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.

[167] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[168] K. El Asnaoui and Y. Chawki, "Using x-ray images and deep learning for automated detection of coronavirus disease," *Journal of Biomolecular Structure and Dynamics*, pp. 1–12, 2020.

[169] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang, *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[170] A. K. Mishra, S. K. Das, P. Roy, and S. Bandyopadhyay, "Identifying covid19 from chest ct images: A deep convolutional neural networks based approach," *Journal of Healthcare Engineering*, vol. 2020, 2020.

[171] M. A. Khan, S. Kadry, Y.-D. Zhang, T. Akram, M. Sharif, A. Rehman, and T. Saba, "Prediction of covid-19-pneumonia based on selected deep features and one class kernel extreme learning machine," *Computers & Electrical Engineering*, vol. 90, p. 106 960, 2021.

[172] H. Chao, X. Fang, J. Zhang, F. Homayounieh, C. D. Arru, S. R. Digumarthy, R. Babaei, H. K. Mobin, I. Mohseni, L. Saba, *et al.*, "Integrative analysis for covid-19 patient outcome prediction," *Medical Image Analysis*, vol. 67, p. 101 844, 2021.

[173] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, pp. 1–13, 2020.

[174] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3619–3629, 2020.

[175] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[176] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, E359–E386, 2015.

[177] M. De Santis, J. Bellmunt, G. Mead, J. M. Kerst, M. Leahy, P. Maroto, T. Gil, S. Marreaud, G. Daugaard, I. Skoneczna, *et al.*, "Randomized phase ii/iii trial assessing gemcitabine/carboplatin and methotrexate/carboplatin/vinblastine in patients with advanced urothelial cancer who are unfit for cisplatin-based chemotherapy: Eortc study 30986," *Journal of clinical oncology*, vol. 30, no. 2, p. 191, 2012.

[178] T. Powles, I. Durán, M. S. Van Der Heijden, Y. Loriot, N. J. Vogelzang, U. De Giorgi, S. Oudard, M. M. Retz, D. Castellano, A. Bamias, *et al.*, "Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (imvigor211): A multicentre, open-label, phase 3 randomised controlled trial," *The Lancet*, vol. 391, no. 10122, pp. 748–757, 2018.

[179] J. Bellmunt, R. De Wit, D. J. Vaughn, Y. Fradet, J.-L. Lee, L. Fong, N. J. Vogelzang, M. A. Climent, D. P. Petrylak, T. K. Choueiri, *et al.*, "Pembrolizumab as second-line therapy for advanced urothelial carcinoma," *New England Journal of Medicine*, vol. 376, no. 11, pp. 1015–1026, 2017.

[180] P. Sharma, P. Bono, J. W. Kim, P. Spiliopoulou, E. Calvo, R. N. Pillai, P. A. Ott, F. G. De Braud, M. A. Morse, D. T. Le, *et al.*, *Efficacy and safety of nivolumab monotherapy in metastatic urothelial cancer (muc): Results from the phase i/ii checkmate 032 study.* 2016.

[181] C. Massard, M. S. Gordon, S. Sharma, S. Rafii, Z. A. Wainberg, J. Luke, T. J. Curiel, G. Colon-Otero, O. Hamid, R. E. Sanborn, *et al.*, "Safety and efficacy of durvalumab (medi4736), an anti–programmed cell death ligand-1 immune checkpoint inhibitor, in patients with advanced urothelial bladder cancer," *Journal of Clinical Oncology*, vol. 34, no. 26, p. 3119, 2016.

[182] A. B. Apolo, J. R. Infante, A. Balmanoukian, M. R. Patel, D. Wang, K. Kelly, A. E. Mega, C. D. Britten, A. Ravaud, A. C. Mita, *et al.*, "Avelumab, an anti–programmed death-ligand 1 antibody, in patients with refractory metastatic urothelial carcinoma: Results from a multicenter, phase ib study," *Journal of Clinical Oncology*, vol. 35, no. 19, p. 2117, 2017.

[183] D. H. Aggen and C. G. Drake, "Biomarkers for immunotherapy in bladder cancer: A moving target," *Journal for immunotherapy of cancer*, vol. 5, no. 1, p. 94, 2017.

[184] S. Paratore, G. L. Banna, M. D'Arrigo, S. Saita, R. Iemmolo, L. Lucenti, D. Bellia, H. Lipari, C. Buscarino, R. Cunsolo, *et al.*, "Cxcr4 and cxcl12 immunoreactivities differentiate primary non-small-cell lung cancer with or without brain metastases," *Cancer Biomarkers*, vol. 10, no. 2, pp. 79–89, 2012.

[185] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv e-prints*, arXiv–1712, 2017.

[186] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[187] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 618–626.

[188] F. Rundo, C. Spampinato, G. L. Banna, and S. Conoci, "Advanced deep learning embedded motion radiomics pipeline for predicting anti-pd-1/pd-l1 immunotherapy response in the treatment of bladder cancer: Preliminary results," *Electronics*, vol. 8, no. 10, p. 1134, 2019.

[189] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, *et al.*, "New response evaluation criteria in solid tumours: Revised recist guideline (version 1.1)," *European journal of cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[190] F. Rundo, G. L. Banna, L. Prezzavento, F. Trenta, S. Conoci, and S. Battiato, "3d non-local neural network: A non-invasive biomarker for immunotherapy treatment outcome prediction. case-study: Metastatic urothelial carcinoma," *Journal of Imaging*, vol. 6, no. 12, p. 133, 2020.

[191] K. H. Cha, L. Hadjiiski, H.-P. Chan, A. Z. Weizer, A. Alva, R. H. Cohan, E. M. Caoili, C. Paramagul, and R. K. Samala, "Bladder cancer treatment response assessment in ct using radiomics with deep-learning," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[192] F. Rundo, G. L. Banna, F. Trenta, C. Spampinato, L. Bidaut, X. Ye, S. Kollias, and S. Battiato, "Advanced non-linear generative model with a deep classifier for immunotherapy outcome prediction: A bladder cancer case study," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, Springer International Publishing, 2021, pp. 227–242.

[193] H. O. Alsaab, S. Sau, R. Alzhrani, K. Tatiparti, K. Bhise, S. K. Kashaw, and A. K. Iyer, "Pd-1 and pd-l1 checkpoint signaling inhibition for cancer immunotherapy: Mechanism, combinations, and clinical outcome," *Frontiers in pharmacology*, vol. 8, p. 561, 2017.

[194] K. R. Spencer, J. Wang, A. W. Silk, S. Ganesan, H. L. Kaufman, and J. M. Mehnert, "Biomarkers for immunotherapy: Current developments and challenges," *American Society of Clinical Oncology educational book*, vol. 36, e493–e503, 2016.

[195] X. Ding, Q. Chen, Z. Yang, J. Li, H. Zhan, N. Lu, M. Chen, Y. Yang, J. Wang, and D. Yang, "Clinicopathological and prognostic value of pd-l1 in urothelial carcinoma: A meta-analysis," *Cancer management and research*, vol. 11, p. 4171, 2019.

[196] T. C. Zhou, A. I. Sankin, S. A. Porcelli, D. S. Perlin, M. P. Schoenberg, and X. Zang, "A review of the pd-1/pd-l1 checkpoint in bladder cancer: From mediator of immune escape to target for treatment," in *Urologic Oncology: Seminars and Original Investigations*, Elsevier, vol. 35, 2017, pp. 14–20.

[197] L. O. Chua and L. Yang, "Cellular neural networks: Theory," *IEEE Transactions on circuits and systems*, vol. 35, no. 10, pp. 1257–1272, 1988.

[198] P. Arena, S. Baglio, L. Fortuna, and G. Manganaro, "Dynamics of state controlled cnns," in *1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96*, IEEE, vol. 3, 1996, pp. 56–59.

[199] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, *Learning transferable architectures for scalable image recognition*, 2017. arXiv: 1707.07012 [cs.CV].

[200] F. Rundo, C. Spampinato, G. L. Banna, and S. Conoci, "Advanced deep learning embedded motion radiomics pipeline for predicting anti-pd-1/pd-l1 immunotherapy response in the treatment of bladder cancer: Preliminary results," *Electronics*, vol. 8, no. 10, p. 1134, 2019.

[201] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, pp. 1–48, 2019.

[202] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, "Multi-task transfer learning deep convolutional neural network: Application to computer-aided diagnosis of breast cancer on mammograms," *Physics in Medicine & Biology*, vol. 62, no. 23, p. 8894, 2017.

[203] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Computer methods and programs in biomedicine*, vol. 187, p. 104 964, 2020.

[204] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[205] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*, Springer, 2017, pp. 146–157.

[206] S. Z. B. Rajemi, R. Hamid, F. Naim, and N. W. Arshad, "Pedestrian detection for automotive night vision using thermal camera," in *2012 7th International Conference on Computing and Convergence Technology (ICCCT)*, IEEE, 2012, pp. 694–698.

[207] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[208] H. Ribeiro, S. Morales, C. Salmerón, A. Cruz, L. Calado, M. I. Rodríguez-García, J. D. Alché, and I. Abreu, "Analysis of the pollen allergen content of twelve olive cultivars grown in Portugal," *Aerobiologia*, vol. 29, no. 4, pp. 513–521, 2013, ISSN: 03935965. DOI: 10.1007/s10453-013-9300-8.

[209] A. Fernstrom and M. Goldblatt, "Aerobiology and Its Role in the Transmission of Infectious Diseases," *Journal of Pathogens*, vol. 2013, pp. 1–13, 2013, ISSN: 2090-3057. DOI: 10.1155/2013/493960.

[210] "Innovations in air sampling to detect plant pathogens," *Annals of Applied Biology*, vol. 166, no. 1, pp. 4–17, 2015.

[211] "Better prediction of Mediterranean olive production using pollen-based models," *Agronomy for Sustainable Development*, vol. 34, no. 3, pp. 685–694, 2014.

[212] M. Cunha, H. Ribeiro, and I. Abreu, "Pollen-based predictive modelling of wine production: Application to an arid region," *European Journal of Agronomy*, vol. 73, pp. 42–54, 2016.

[213] "Principles and methods for automated palynology," *New Phytol*, no. 1996, pp. 1–8, 2014, ISSN: 0028646X. DOI: 10.1111/nph.12848.

[214] J. Oteros, G. Pusch, I. Weichenmeier, U. Heimann, R. Möller, S. Röseler, C. Traidl-Hoffmann, C. Schmidt-Weber, and J. T. Buters, "Automatic and online pollen monitoring," *International Archives of Allergy and Immunology*, vol. 167, no. 3, pp. 158–166, 2015, ISSN: 14230097. DOI: 10.1159/000436968.

[215] S. Kawashima, M. Thibaudon, S. Matsuda, T. Fujita, N. Lemonis, B. Clot, and G. Oliver, "Automated pollen monitoring system using laser optics for observing seasonal changes in the concentration of total airborne pollen," *Aerobiologia*, vol. 33, no. 3, pp. 351–362, 2017, ISSN: 15733025. DOI: 10.1007/s10453-017-9474-6.

[216] "Pollen and spore monitoring in the world," *Clinical and Translational Allergy*, vol. 8, no. 1, pp. 1–5, 2018.

[217] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[218] C. M. Bishop, *Pattern Recognition and Machine learning*. Springer, 2006.

[219] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[220] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 886–893.

[221] N. A. Chinchor and B. Sundheim, "Message understanding conference (muc) tests of discourse processing," in *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995, pp. 21–26.

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

[223] R. Du and et al., "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *European Conference on Computer Vision*, Springer, 2020, pp. 153–168.

[224] S. Battiato, A. Ortis, F. Trenta, L. Ascari, M. Politi, and C. Siniscalco, "Pollen13k: A large scale microscope pollen grain image dataset," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 2456–2460.

[225] S. Battiato and et al., "Pollen grain classification challenge 2020," in *Pattern Recognition. ICPR International Workshops and Challenges*, Cham: Springer International Publishing, 2021, pp. 469–479, ISBN: 978-3-030-68793-9.

[226] P. Gui and et al., "Improved data augmentation of deep convolutional neural network for pollen grains classification," in *Pattern Recognition. ICPR International Workshops and Challenges*, Cham: Springer International Publishing, 2021, pp. 490–500, ISBN: 978-3-030-68793-9.

[227] C. Wei and et al., "Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1910–1919.

[228] M. Buda and et al., "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[229] S. Yun and et al., "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[230] C. Fang and et al., "The fusion of neural architecture search and destruction and construction learning," in *Pattern Recognition. ICPR International Workshops and Challenges*, Cham: Springer International Publishing, 2021, pp. 480–489, ISBN: 978-3-030-68793-9.

[231] Y. Chen and et al., "Destruction and construction learning for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5157–5166.

[232] J. Fang and et al., "Densely connected search space for more flexible neural architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 628–10 637.

[233] S. Battiato, A. Ortis, F. Trenta, L. Ascari, M. Politi, and C. Siniscalco, "Detection and classification of pollen grain microscope images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 980–981.

[234] J. D. Hamilton, *Time series analysis*. Princeton university press, 2020.

[235] R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, J. P. Nobrega, and A. L. Oliveira, "Computational intelligence and financial markets: A survey and future directions," *Expert Systems with Applications*, vol. 55, pp. 194–211, 2016.

[236] F. Rundo, F. Trenta, A. L. di Stallo, and S. Battiato, "Machine learning for quantitative finance applications: A survey," *Applied Sciences*, vol. 9, no. 24, p. 5574, 2019.