Dottorato di Ricerca in Informatica
Ciclo 34

Ph.D. Thesis

Artificial Intelligence methods for supporting biomedical data analysis

Author: **Maria Frasca**

Supervisors: **Prof. Michele RISI**

Prof. Rita FRANCESE

PhD Program Director: **Prof. Andrea DE LUCIA**

A.A 2020/2021

Abstract

Artificial intelligence (AI) is a fundamental technology useful in many fields, ranging from finance, weather forecast, to medicine. Positive results are reached especially in the diagnostic field, where AI provides a great support to the physician's assessments. The goal of this thesis is to support biomedical data analysis by means Artificial Intelligence methods and techniques. To this aim I define processes based on Machine Learning capable of improving diagnoses, by identifying the schematics of a well-determined pathology. These tools will be available to clinicians, in order to be able to intervene on patients through countermeasures adapted to their specific needs. I have concentrated my attention on the detection two kinds of diseases: (i) the degenerative disease, such as Parkinson and Coloboma, and (ii) the oncological disease, such as Leukemia and Melanoma.

In particular, in the case of Parkinson, it is difficult to formulate a clinical diagnosis because there are neither objective tests nor specific biochemical and neuroradiological markers. I apply IR techniques to patient records belonging to a standard dataset for classifying Parkison patients on the base of the reports produced during the different visits. The obtained results are very promising on the use of this tool in the clinical practice. I also investigate how traditional biometric techniques may fail in presence of a iris pathology, such as Coloboma. Thus, I adopted Artificial Intelligence techniques for detecting irises affected by Coloboma, I demonstrated that traditional biometrics alghorithms fail in presence of this disease, such as the ones proposed by Daugman and Canny. Thus, I develop an algorithm which extends the largely adopted Daugman's algorithm and allows also the people affected by this disease to be recognized by biometrical systems. In this way, they are not excluded by the access a services secured by iris detection. I also experiment Artificial Intelligence models and technique to detect in the case of the detection of oncological disease. In particular, concerning Leukemia I define a process aiming at detecting a set of differentially expressed genes in terms of methylation level, i.e., genes that in different conditions have an expression level significantly different in the Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) cases, and their characteristic pathways. The detection of gene expression data samples involves feature selection and classification. To this aim, Deep Learning models have been adopted (e.g., feature selection techniques and classifiers methods). A methodology is also proposed for the classification of melanoma by adopting different Deep Learning techniques applied to a common image dataset extracted from the ISIC dataset and consisting of different types of skin diseases, including melanoma on which is applied a specific pre-processing phase. The results of the adopted techniques (i.e., ResNet, 2D CNN, and SOM) are compared to select the best effective neural network for the recognition and classification of melanoma and evaluate the impact of the pre-processing phase. I also propose an augmented reality applica-

tion for to support of the diagnosis of melanoma. It exploits both Artificial intelligence and image processing techniques. I describe in detail the real-time process proposed to display the augmented nevus information and evaluated the real-time performances and the app usability. The main results of the proposed approaches are encouraging and suggest that they may be considered in the practical clinical. In the future, I plan to use artificial intelligence models and techniques and image processing to analyze Magnetic Risonance Images of the brain to detect progression of Parkinson's disease. Furthermore, the results obtained in the case of Coloboma of the eye could be extended and studied in other ophthalmic diseases.

Contents

List of figures **9**

List of tables **11**

1 INTRODUCTION **1**
1.1 Motivation . . . 1
1.2 Research statement . . . 3
1.3 Research contribution . . . 3
1.4 Structure of the thesis . . . 6

2 BACKGROUND **9**
2.1 Artificial Intelligence . . . 9
2.1.1 History of Artificial Intelligence . . . 11
2.2 Bioinformatics . . . 13
2.3 Iris recognition . . . 15
2.3.1 Daugman's algorithm . . . 15
2.3.2 Canny's algorithm . . . 17

3 RELATED WORK **19**
3.1 AI support to degenerative diseases . . . 19
3.1.1 Related on Text Mining e Information Retrieval Tecnique . . . 19
3.1.2 Related on Iris ricognition in patological state and descrimination in IoBT service . . . 21
3.2 AI support to oncological diseases . . . 23
3.2.1 Melanoma detection . . . 23
3.2.2 AI for feature selection of deferentially expressed genes . . . 26

4 ARTIFICIAL INTELLIGENCE IN DEGENERATIVE DISEASES **29**
4.1 Supporting Parkinson's detection . . . 29
4.1.1 Parkinson's disease . . . 30

4.1.2 Dataset . . . 31
4.1.3 The proposed process . . . 33
4.2 AI for supporting the detection of Coloboma Disease . . . 46
4.2.1 What is Coloboma? . . . 46
4.2.2 Dataset . . . 46
4.2.3 On the Limitation of Pathological Iris Recognition: Neural Network Perspectives . . . 48
4.2.4 Extension of the Daugman algorithm . . . 59

5 A.I. IN ONCOLOGICAL DISEASES **83**
5.1 Supporting Melanoma detection . . . 84
5.1.1 Melanoma . . . 84
5.1.2 Dataset . . . 86
5.1.3 A Comparison of Neural Network Approaches for Melanoma Classification . . . 87
5.1.4 An Augmented Reality Mobile Application for Skin Lesion Data Visualization . . . 98
5.2 Leukemia classification . . . 113
5.2.1 Leukemia . . . 114
5.2.2 Dataset . . . 116
5.2.3 The proposed approach . . . 117

6 DISCUSSION AND CONCLUSION **131**
6.0.1 International Journals . . . 133
6.0.2 International Conferences . . . 133

APPENDIX A: THRESHOLD SETTING **157**

List of Figures

4.1 The data model of the PPMI dataset. 32
4.2 Graphical representation of the clusters obtained with K-means (LSA (a), Text2Vec (c), Doc2Vec (e)) and Fuzzy c-means (LSA (b), Text2Vec (d), Doc2Vec (f)) algorithms for each technique. 38
4.3 Clustering results for the K-means algorithm and Precision, Recall and F-measure data. 40
4.4 Clustering results for Fuzzy c-means algorithm and Precision, Recall and F-measure data. 41
4.5 Comparison among F-measures for k-means and Fuzzy c-means algorithms. 43
4.6 Barplot representing the words with the highest frequencies in the documents related to the LOG visits. 44
4.7 A data visualization based or Radar chart for fast diagnosis. 45
4.8 An eye suffering from Coloboma disease. 49
4.9 Visualization of errors made by Daugman's algorithm (left) versus true pupils and iris edges (right). 50
4.10 Visualization of errors made by Canny edge detection algorithm (left) versus true pupils and iris edges (right). 52
4.11 Error values between (a) the Daugman output ((b) the Canny output) and the true edges of iris and pupil. 53
4.12 CNN to classify iris suffering from Coloboma. The CNN is structured in: 4 *Convolutional blocks*, 1 *Flattering layer*, and 1 *Dense layer*. . . . 54
4.13 Loss and Accuracy curves over the training epochs. 56
4.14 ResNet to classify iris suffering from Coloboma. 56
4.15 Accuracy (a) and Loss (b) curves over the training epochs. 59
4.16 Process of creating masks. 60
4.17 Boundaries ed Intersections. 61
4.18 Resize down with scaling 1.5 62
4.19 Overall pupil scaling process 64
4.20 Complete Segmentation Process. 65

4.21 Localization of the iris. 67
4.22 Localization of the Pupil. 69
4.23 Localization of the Pupil. 71
4.24 Normalization process. 72
4.25 Errors in the implementation of circle and ellipse. 74
4.26 Comparison of the overall errors of the two implementations. 75
4.27 Comparison between the average errors in the two implementations. 76
4.28 Detection example 76
4.29 Formula to find the error 77
4.30 Representation of the iterative process of matching between images with the same scaleOverall. 79
4.31 Representation of the iterative matching process at the change of scaleOverall. 80

5.1 The ABCDE rules. 86
5.2 Hair removal by using the Canny edge detector. 88
5.3 Process for the lesion segmentation and extraction. 90
5.4 The Convolutional Neural Network architecture. 91
5.5 The full implementation of the Residual Neural Network. 93
5.6 The main phases of SUSI. 94
5.7 U-matrix: visual representation of the distances between neurons in the SOM network. 97
5.8 The ROC curve of the ResNet model. 98
5.9 The main visualization layout. 102
5.10 Select and center a nevi (a), the AR visualization (b). 103
5.11 The real-time process. 104
5.12 The skin lesion image processing. 106
5.13 The Convolutional Neural Network architecture. 107
5.14 Nevus tracking. 108
5.15 Average time for task (a) and number of errors for task (b). 112
5.16 The Post-Experiment questionnaire results. 113
5.17 Acute Lymphoblastic Leukemia (a) and Acute Myeloid Leukemia (b). 115
5.18 Visualization of the process in microarray analysis [234]. 117
5.19 The data analysis process based on Bayesian and Autoencoders feature selection. 118
5.20 The autoencoder architecture, where n represents the number of genes. 123
5.21 The DNN architecture, where n represents the number of genes. . . 123

5.22 Loss and accuracy results of the applied neural network on the feature selection implemented with Limma and autoencoders, where "val"/"$loss$" represents the average accuracy/loss of the training set, and "val_acc"/"val_loss" represents the average accuracy/loss of the validation set, respectively. 124
5.23 The data analysis process based on Genetic Algorithm feature selection. 124
5.24 Steps of the GA for feature selection. 125
5.25 The "RNA degradation" pathway [216]. 128
5.26 Gene enriched from "RNA degradation" pathway. 129

6.1 Probability of false matches with increasing HD threshold. 158
6.2 Results of intra-class and inter-class matching on non-ideal images. 159

List of Tables

3.1 Comparison with different neural networks on different dataset. . . 25
3.2 Features selection approaches applied on microarray data. 27

4.1 The dataset composition 47
4.2 Iris caption 73
4.3 Pupil caption 77
4.4 Overall surveys 77
4.5 Results obtained for the entire Dataset 81

5.1 The neural networks results **without** pre-preprocessing. 96
5.2 The neural networks results **with** the pre-processing. 96
5.3 The CNN classification results. 107
5.4 The tasks composing the scenario of use. 109
5.5 Post-Experiment questionnaire. 109
5.6 Average process performance measures. 114
5.7 GA parameters. 126
5.8 Comparison with the adopted features selection approaches. 126

6.1 Datasets used by Libor Masek for testing 157
6.2 False Acceptance and False Rejection Rate for dataset CASIA-a . . . 160
6.3 False Acceptance e Rejection Rate for dataset LEI-a 160

Chapter 1

INTRODUCTION

1.1 Motivation

Artificial intelligence is widely used in medicine and bioinformatics research. Bioinformatics combines biology and information system to examine the procedure and classify biological data in a short time.

Various artificial intelligence algorithms have been developed and used in bioinformatics analyses. They play an important role in this field to rationalize complex systems and to perform multidisciplinary analysis, by enabling pattern recognition in complex biological data.

Artificial intelligence has been used in medicine since the 1950s to improve diagnoses using computer-assisted programs [31, 241].

Interest and advances in AI applications in medicine have increased in recent years due to the growing computing power of modern computers and a large amount of digital data available for their collection and use [155]. There are several applications of artificial intelligence that can be used in a variety of medical fields, such as clinical, diagnostic, rehabilitative, surgical, predictive practices. Artificial intelligence technologies can acquire, analyze and report large volumes of data in a variety of ways to detect disease and guide clinical decisions [44]

The "machine learning" algorithms operate on large data sets, such as a considerable amount of clinical cases to which a diagnosis or prognosis has been associated. Based on the observation of the incoming and outgoing relationships they are able to learn statistical models that can be used to make predictions in the presence of new data.

One of the areas that have given a strong boost to artificial intelligence is represented by "deep learning" techniques, the advantage of which is the ability to directly analyze information that is typically not used in statistical analysis, such as

images, written texts in natural language or data from wearable devices. They are tools that allow us to analyze a lot of data with great effectiveness, improving our speed of information interpretation. In the oncology field, there are numerous applications of "deep learning" especially for image analysis in many fields, such as pathological anatomy and radiology. As an example, in the case of magnetic resonances the disease progression analysis can be automated using a combination of classical analysis and deep learning techniques.

According to research by Frost & Sullivan, the market for AI in healthcare will reach 6 billion dollars in 2022, with an annual growth rate of 68%, generating savings of over 150 billion dollars. Global Market Insights estimates that there will be an annual growth of 41.7% until 2025. In the medical imaging sector alone, one of the most promising among those already available, the forecast is that the AI market will record a growth of 30% per year in the period up to 2025, thanks to the improvement in computing power, learning algorithms and the availability of ever-larger data sets.

Very important is the application of AI in predictive diagnostics: through the use and interpretation of data, the first signs of some diseases can be grasped to help clinicians make more accurate diagnoses, with the aim of reducing errors and developing methods for individualized medical treatment. For this reason, we speak of computer-assisted diagnosis (CAD), which are systems that assist clinicians in the interpretation of medical images. CAD is an interdisciplinary technology that combines elements of artificial intelligence and artificial vision with the processing of radiological and pathological images. A typical application is the detection of a tumor.

In the late 1950s, with the advent of modern computers, researchers in various fields began to explore the possibility of building computer-aided medical diagnostics (CAD) systems. These early CAD systems used flowcharts, statistical pattern-matching, probability theory, or knowledge bases to guide their decision-making.

Despite the many developments that CAD systems has achieved since the dawn of computers, there are still some challenges that these systems face today. Some challenges are related to various algorithmic limitations in the procedures of a CAD system, including input data collection, preprocessing, processing, and system evaluations. Algorithms are generally designed to select a single probable diagnosis, thus providing suboptimal results for patients with multiple and concomitant disorders. Today, the input data for CAD comes mainly from electronic health records (EHR). Effective design, implementation, and analysis for EHR is a fundamental necessity for any CAD system.

Due to the massive availability of data and the need to analyze that data, big data is also one of the biggest challenges facing CAD systems today. The increasing amount of patient data is a serious problem. Patient data is often complex and can be semi-structured or unstructured data. It requires highly developed approaches to archive, retrieve and analyze them in a reasonable time.

During the preprocessing phase, the input data must be normalized. Normalizing the input data includes noise reduction and filtering. Processing may contain some minor steps depending on the applications. The three basic steps of medical imaging are segmentation, feature extraction/selection, and classification. These sub-steps require advanced techniques to analyze the input data with less computation time. Although much effort has been devoted to creating innovative techniques for these CAD systems procedures, there is still no best algorithm for each step. Continuous studies are essential in the construction of innovative algorithms for all aspects of CAD systems. There is also a lack of standardized evaluation measures for CAD systems.

At the present, in many cases these tools cannot replace clinicians yet: these technologies provide assistance, helping healthcare professionals to grasp significant elements that would otherwise remain hidden, extracting them from huge amounts of data.

1.2 Research statement

The applications of AI in the medical field are many and still need to be explored. In this direction, this thesis aims at answering the following main research question:

RQ: How Artificial Intelligence may support biomedical data analysis?

To answer the main research question RQ, which is very wide, we experimented the adoption of AI methods and techniques in several medical areas.

In particular, we addressed the following sub-research questions:

- **RQ1:** How Artificial Intelligence may support the degenerative disease identification?
- **RQ2:** How Artificial Intelligence may support the oncological disease classification?

1.3 Research contribution

The contribution of this thesis is composed of two parts: the support AI offers to the detection of degenerative desease (RQ1) and to the detection of oncological desease (RQ2).

- **Support to degenerative disease (RQ1).** We considered two degenerative disease as case study: Parkinson and Coloboma. Concerning the former, the ba-

sic idea is to analyze the patient medical records and identify eventual correlations between documents to recognize the different classes of patients. To this aim, a technique has been proposed to identify a correlation between the biomedical data in the PPMI (Parkinson's Progression Markers Initiative) dataset for verifying the consistency of medical reports formulated during the sequence of visits and allow to correctly categorize the various patients. To correlate the information of each patientâs medical report, IR and ML techniques have been adopted, including the Latent Semantic Analysis (LSA), Text2Vec and Doc2Vec approaches. Then, patients are grouped and classified into affected or not by using clustering algorithms (i.e., K-means and Fuzzy clustering) according to the similarity of medical reports. Finally, a visualization system has been adopted based on the D3 (Data-Driven Documents) framework to visualize correlations among medical reports with an interactive chart, and to support the clinician in analysing the chronological sequence of visits in order to diagnose Parkinson's disease early. From what has been observed, the processes with LSA produce the worst results, and generally, the results produced by the execution of the processes via K-means are lower than those with Fuzzy clustering. This result concerns the intrinsic structure of the K-means algorithm that does not allow an element that can be positioned in both clusters or that it can be moved later from one cluster to another, forcing an incorrect classification of the information processed. This is because in our data there is the PRODROMAL class, which are patients not affected by Parkinson's disease but who present symptoms characteristic of this disease. As a consequence, a stringent clustering could lead to a higher error rate, as can be seen from the results. In conclusion, the techniques that currently produce the best results are: *i)* Text2Vec with Fuzzy, in visits from SC (screening) to V02 and from V10 to LOG, and *ii)* Doc2VecM with Fuzzy in visits from V03 to V09. In the future, the analysis of the data and the correlation between them may be further extended.

Coloboma is a congenital abnormality of membranes of the eye. The research investigated the difficulty of iris recognition when the eye is affected by this disease. The results have shown how this pathological state impacts the quality of the result of the Daugman and Canny edge detection algorithms, which represent the most widespread methods used for the iris localization step in eye-based biometric. From our results it emerged that the presence of Coloboma seems not to be appropriate for iris recognition. This disease may become a sort of biometric-driven discrimination, which may be very detrimental in an environment riches of IoBT (Internet of Biometric Things) services. Daugman and Canny algorithms have been executed on a dataset of irises with presence of Coloboma obtaining a significant error rate for both the algorithms, i.e., 71.05% of analysed irises for Daugman and 52.63% for Canny. Thus, by consid-

ering the high error in the recognition of the iris, the risk of exclusion for people suffering from Coloboma is high and concrete. To improve the classification of the state of the iris in Coloboma and in non-pathological cases, an approach based on ResNet has been shown proposed that archieves 99.79% accuracy. This is an encouraging result towards a deeper analysis addressing many aspects, such as other ophthalmic disease states and the use of larger datasets involving different Coloboma stages. For people who reveal Coloboma malformation there is the need of creation of "diversity-aware" biometric tools which detects the anomaly and proposes an alternative authentication approach. In addition, an extension of the Daugman algorithm has been specifically designed for the detection of the characteristics of eyes affected by pupil malformations, due to diseases such as coloboma, for which the Daugman algorithm does not guarantee efficient results. For this reason there is a need to go beyond the original assumption for which the iris and pupil can be described by two concentric circumferences. This extension is mainly focused on the first segmentation phase, in which the pupil is identified through an elliptical shape and is based on the implementation of Libor Masek, used by the principle for the detection of the Daugman algorithm. Thus, obtaining an errors of 8% and 22% in the detection of the pupil and iris respectively compared to the 22% and 50% obtained from the implementation of the algorithm of Libor Masek, and a total error of 40% compared to 72%. Overall, there was an improvement of 15% on the average error and 57% on the average of the maximum errors.

- **Support to oncological diseases (RQ2).** We focused our attention on two disease: Leukemia and Melanoma. Concerning Leukemia, we analyzed two different types of Leukemia: Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). The aim is to detect a set of differentially expressed genes in terms of methylation level, i.e., genes that in different conditions have an expression level significantly different in the AML and ALL cases, and their characteristic pathways. To this purpose, a Data Science process has been defined. The detection of gene expression data samples involves feature selection and classification. In particular, Deep Learning models have been adopted (e.g., feature selection techniques and classifiers methods). The analysis has been performed on a dataset consisting of samples from people with Leukemia, characterized by a fixed list of genes; the samples belong to two distinct classes: ALL and AML. Following the reduction in the number of variables, classification models have been implemented with the use of a Deep Neural network (DNN), obtaining a classification accuracy of approximately 92%. Then, the results have been compared with the ones provided by an approach based on Support Vector Machines (SVM) giving an accuracy of 87,39%. Moreover, another feature selection approach based on genetic algorithms has been exper-

imented obtaining 60,36% (DNN) and 30,63% (SVM) for accuracy. For further verification of the relevance of the selected set of genes, we conducted a gene enrichment analysis based on the functional annotation of the differentially expressed genes. A relevant result is the detection of a differentially expressed pathway between the two pathologies.

The other oncological disease investigated in this thesis is melanoma. A methodology has been proposed for the classification of melanoma by adopting different Deep Learning techniques applied to a common image dataset extracted from the ISIC dataset and consisting of different types of skin diseases, including melanoma on which has been applied a specific pre-processing phase. The results of the adopted techniques (i.e., ResNet, 2D CNN, and SOM) have been compared to select the best effective neural network for the recognition and classification of melanoma and evaluate the impact of the pre-processing phase. The better result was obtained with the ResNet (i.e., 81,5% for accuracy) Finally, our results have been compared, in terms of sensitivity and specificity, with a study carried out chosen from 12,000 dermoscopic images in the ISIC archive. It is worth noting that the manual approach performed by dermatologists achieved a mean sensitivity of 82% with a mean specificity of 59%. From these results, our analysis has obtained a better accuracy than dermatologists with the ResNet, a greater sensitivity with the 2D CNN and the ResNet and a greater specificity with all the adopted neural networks. I also devised an application for the recognition of the lesion basing its reality increased as a tool to support the measurement of the diagnosis of melanoma. I described in detail the real-time process proposed to display the augmented nevus information and evaluated the real-time performances and the app usability. Because of the pandemic period conducting empirical work (e.g., recruiting a large number of participants) was challenging.

1.4 Structure of the thesis

The reminder of this thesis is structured as follows.

- **The second chapter** presents an overview of AI and discusses how important its application in the field of Bioinformatics is.
- In the **third chapter** I discuss related work concerning the application of AI methods and techniques to the detection of both degenerative and oncological diseases.
- in **fourth chapter** I describe the support AI may provide in the diagnosis of neurodegenerative diseases. Specifically, I will deal with Parkinson's disease and Coloboma of the eye.

- In the **fifth chapter**, I will investigate how AI can support the diagnosis of oncological diseases, such as Melanoma and Leukemia.
- The **sixth chapter** concludes the thesis with final remarks and future work.
- A discussion on the setting of the biometric threshold adopted in the extension of the Daugman's algorithms is reported in the **Appendix A**

Chapter 2

BACKGROUND

In this section I introduce the main concepts on which provide the basis for this thesis work.

2.1 Artificial Intelligence

Artificial Intelligence is defined as a series of studies and techniques typical of information technologies, the purpose of which is to create technological programs and systems capable of solving problems and performing tasks that are normally understood by the human mind and body attributed to capabilities human [210]. With recent advances, it is possible to identify Artificial Intelligence (hereinafter "A.I.") as the discipline, that deals with allowing machines (hardware and software) to function autonomously [230]. According to a strictly computerized meaning, the A.I. could be classified as the discipline that encompasses the theories and practical techniques for the development of algorithms that allow machines (especially computers) to show intelligent activity, mostly in specific domains and application fields. This definition requires a formal classification of the synthetic/abstract functions of human reasoning, meta-reasoning, and learning; in relation to this observation, the issue of 'Cognitive Computing' emerges, understood as the set of technological platforms based on the scientific disciplines of AI. (including $Machine\ Learning$ and $Deep\ Learning$) and $Signal\ Processing$ (i.e. the ability to process signals). The increasing attention paid to this discipline is motivated by the results obtained thanks to the technological maturity achieved both in computer-assisted computing and in the ability to analyze large amounts of data in any form in real-time and in a short period. ($Big\ data\ analysis$). AI is widespread in the everyday life of most people: the various voice recognition tools that are used regularly (from smartphones to security systems) are based on typical AI algorithms, in the market values, in medicine (use

of neural networks in the analysis of the heartbeat, in the diagnosis of some types of cancer [186], and the development of companion robots [66]), and robotics.

Definitions of A.I. can be divided into two categories: those that relate to thought processes that go back to specific human activities, and those that describe behavior, distinguishing between a $functional$ and a $structural$ approach. The functional approach, also called $behaviorist$, conceives intelligence independently of the physical structure of the computer that implements the intelligent system and aims only at emulation (also selectively). On the contrary, according to the $structural$ approach, also called $constructivist$ or $connectivist$, intelligence is obtained by simulating the human brain and reproducing its structure and properties (connectionist architectures and neural networks). From an operational point of view, two different approaches to A.I. can be distinguished: a $top-down$ and a $bottom-up$. The top-down approach is essentially independent of the underlying level (whatever, computer or brain) and typically takes a symbolic approach: mental states are identified with symbolic representations within a physical-symbolic system. The bottom-up or connectionist approach, on the other hand, relies on artificial neuron architectures or networks that simulate brain neurons to build more complex argumentation structures and methods. Symbolic approaches understand thought formally as the result of symbol manipulation, whereas with neural or connectionist approaches, thought is implicitly determined by the interconnection and distributed processing of many simple computing units. It follows that symbolic approaches are more transparent and easier for individuals to interpret, while connectionist approaches more easily face changing, uncertain, incomplete, and dynamic realities at the expense of transparency. Recent developments and great recent results of bottom-up approaches to neural networks increasingly show the effectiveness of these methods, especially in the field of perception.

Observation of nature is a source of inspiration in A.I., an example is the $Swarm\ Intelligence$ which shows the realization of intelligent behaviors from a collective perspective. The (interdisciplinary) study of these phenomena has made it possible to develop intelligent systems based on robust and adaptive natural models and processes and used to solve research problems, optimization planning, data analysis, and robot coordination. An example is $smart$ objects (such as smartphones, cars, drones) in the context of the Internet of Things (IoT). Always starting from the observation of nature and, in particular, from the selection of the species to adapt to the environment, the $genetic\ algorithms$ have been defined, which draw inspiration from the theory of natural evolution, developed by John Holland starting from the '70s. In these algorithms, the search for a solution is based on the identification of a particular winning generation. In a simplified way, starting from an initial configuration, ed evolving on the basis of natural laws, it is possible to create a new generation (which could be interpreted as a solution). The fitness function, capable of evaluating the characteristics of a generation or solution, ensures the selection of

the best solutions in the field of reproduction; mutation ensures the introduction of new elements randomly within generations, while reproduction guarantees the combination of good solutions in a new one. The system evolves from generation to generation up to a generation (or solution) considered satisfactory.

2.1.1 History of Artificial Intelligence

Starting in the 40s, we have witnessed the spread of the term $cybernetics$, understood as a systematic investigation of communication and control processes in living beings and machines; The basic idea is to examine the self-regulation and control mechanisms that exist in both natural and artificial organisms with feedback capacity; Among the first results of the scientific community in this area is the 1943 project led by researchers Warren McCulloch and Walter Pitt, which proposed a neural network model inspired by the functioning of the human brain; the examined model, which consists of neurons connected through synapses, was able to implement the blocks of Boolean logic. In 1949, Canadian psychologist Donald Olding Hebb proposed a combined study of data from the physiology of the nervous system and the analysis of human behavior, thanks to which the connections between artificial neurons and complex models of the human brain were analyzed in detail. This shows that a modification of the connecting forces between neurons could trigger learning processes. In 1951, the American mathematician and scientist Marvin Lee Minsky created the first neural network-based computer called $SNARC$ ($Stochastic\ Neural\ Analog\ Reinforcement\ Computer$), which could simulate a network of 40 neurons.

The first functional prototypes of neural networks, that is models mathematical/computer models designed to reproduce how biological neurons work to solve A.I. (understood as the ability of a machine to perform functions and arguments like a human mind) emerged in the late 1950s. The concept of $ideal\ machine$ can be found as early as 1936 thanks to the conception, by the English mathematician and cryptographer Alan Mathison Turing, of the so-called $Turing\ machine$, which is an abstract model of a machine capable of executing algorithms and equipped with a potentially infinite tape on which to read and write symbols; The model in question represents a theoretical tool widely used in computability theory and in the study of the complexity of algorithms to understand the limits of mechanical computation. Another significant contribution, attributable to the same scholar, is the so-called $Turing\ Test$ [159], a criterion proposed in 1950 in the article $Computer\ science, machines\ and\ intelligence$ to determine whether a machine can be capable of thinking.

The official event that marks the birth of A.I. is the $Dartmouth\ Summer\ Research\ Project\ on\ Artificial\ Intelligence$, carried out in 1956 at Dartmouth College, which was attended by renowned names in the field of intelligent systems such as the math-

ematician John McCarthy of Dartmouth College (inventor of the term $Artificial\ Intelligence$), the scientist Marvin Lee Minsky of Harvard University, the computer scientist Nathaniel Rochester of the IBM Corporation, the engineer Claude Elwood Shannon of the Bell Telephone Laboratories, with the aim of defining the discipline of AI and developing some research projects to simulate the human intelligence and with it the launch of the first programming languages (Lisp in 1958 and Prolog in 1973) specific for A.I. In the following years, the focus was on the computer and its perception ranged from the arithmetic processor to the high-level machine capable of solving problems and processing symbols.

In 1958 the American psychologist Frank Rosenblatt proposed the $perceptron$, an electronic device with an input and output layer and a learning rule based on the minimization of errors, the function called $Error - Backpropagation$, which is based on the evaluation of the actual output of the network concerning a given input changes the weights of the connections (synapses) as the difference between the actual output and the desired output; In the following years, some scientists demonstrated its limits (the ability, after appropriate training, to recognize only functions in a linearly separable way: a multilevel network of $perceptron$ could have solved more complex problems, but the increasing complexity computational training made the development too onerous for the moment).

The following are the first attempts to develop a human-machine dialogue that can mimic human-human dialogue: One of the best known is the chatbot $Eliza$, software that was developed in 1966 by the German computer scientist Joseph Weizenbaum and based on had aim to simulate a conversation with a human. In addition, over the years other tests have been proposed to highlight other skills considered essential, such as that the intelligent agent not only has a mere capacity for analysis but also a great capacity to interpret and use common sense.The first $expert\ systems$ appeared in the 1970s: A.I. Systems that have been developed to solve a specific task by being modeled on a person who specializes in this area in a limited area and who often also offers explanations of the argumentation mechanisms used. The inference mechanisms used to solve the problem use the knowledge base derived from the knowledge of the domain expert and are implemented by an inference engine that implements the appropriate inference algorithms. The first successes of Expert Systems were achieved in the 1980s with its use in various areas such as diagnostics, design, monitoring, data interpretation, and planning: in the same years, interest in A.I. It is aimed at the industry. Significant research funds are used to fund A.I. projects in the United States, Europe, and Japan; Also in the field of application, neural networks are the subject of renewed interest. Later we realize that expert systems have obvious general limits to solving problems and bottlenecks in their implementation, which mainly result from the difficulty of manually configuring and updating knowledge databases.

In the 90's we see the birth of the World Wide Web and the market entry of graph-

ics processors, which support complex processes much faster, run at lower frequencies, and consume less power than older CPUs. , which leads to a wide and rapid spread of the Internet, allows access to vast amounts of information and knowledge, and opens up new perspectives for artificial intelligence. Opened. through the development of algorithms and applications made possible by the availability of large amounts of unstructured data and the increasing availability of inexpensive computing power. In the last decade, so-called $neuromorphic\ chips$ have also been developed, that is, microchips that integrate data processing and storage in a single micro-component to emulate the sensory and cognitive functions of the human brain.Learning systems and algorithms have become increasingly effective and efficient, with much refinement in the techniques related to neural architectures (even with multiple layers) with incremental and not necessarily supervised learning; Machine learning is used successfully, for example, in document classification and processing, in natural language understanding, in bioinformatics and image processing. Increasingly efficient methods of speech recognition and image classification have been developed and used successfully in robotics and image processing; Many of the web search algorithms, translators, speech recognizers, image classifiers, and photographs that we use daily benefit from these ever-changing techniques. A.I. is capable of supporting new technologies that are evolving rapidly, both in terms of the design of the most appropriate tools and in terms of methodological contribution (for example, in real-time the information generated by those of interest Understand situations automatically and plan actions in dynamic contexts). The use of A.I. The techniques allow a wide range of applications such as integrated monitoring, monitoring, and diagnosis systems, tele-assistance systems and logistics transport planning, autonomous vehicles, etc.; The availability of technological tools for home automation opens the possibility of application to problems related to the aging of the population [99]. Another very interesting field of application is the so-called $Internet\ of\ the\ future$, characterized by being an open network of self-organized and intelligent units such as software (agents, web services, softbots, avatars), hardware (objects, sensors, robots).

2.2 Bioinformatics

Bioinformatics is an interdisciplinary research area that deals with developing new algorithms, methodologies, and software tools for analyzing biological data. Bioinformatics combines computer science, statistics, mathematics, and engineering to study and extract new knowledge from biological data. Historically, bioinformatics was born to use the computer to analyze the sequences of genes and proteins, but the sector has progressively expanded to include the management, processing, analysis, and visualization of large quantities of data produced. from genomic research, proteomics, drug control, and combinatorial chemistry, but also image analysis to

search for characteristic patterns. Bioinformatics, therefore, includes the integration and exploration of increasingly extensive databases of biological interest. However, it is possible to identify three main areas which constitute and characterize this science [189]. They concern:

- The development of applications to support laboratory experiments [62, 131, 151, 169, 200]. These are tools that allow automating sequences reconstruction procedures, typical in the shotgun technique, but also tools that allow managing data with a lower level of detail, such as the positioning of markers on genetic maps [7, 201, 202]. This area also includes all those software that allows you to manage information deriving from data acquisition techniques other than sequences such as microarrays [142, 152, 246] and a radiation hybrid [165, 211].

- The design, implementation, and integration of databases. It is conceivable that the growth trend in the number of databases will also be positive in the future. And this is mainly due to two reasons. The first concerns the fact that the completely sequenced genomes are limited in number compared to those in which the work is in progress, the second comes from the need to note the results of analyzes carried out on the current databases. Examples are those that contain sequences recognized as those active in particular biological functions such as ProDom, Prosite [104], Prints-s [12], and Blocks.

- Techniques and applications for data-mining [25, 50, 78, 95, 96, 133]. This area concerns the second phase of the genome project and is inherent in the extraction of information/knowledge from the data produced [187]. Although it was presented under the single name of data-mining this area is very heterogeneous, including divisions that in common have the sole purpose of extracting knowledge and the fact of addressing similar issues.

Bioinformaticians expect that the project of life emerges from this work; that is, that new models or explanatory principles come to light capable of functionally correlating genes with metabolic pathways, giving an evolutionary meaning to the comparison between genes and gene pools of different species, capturing the logic that governs the three-dimensional assembly of proteins, and then predict the function of genes and proteins by integrating the different types of information available. One of the most important objectives is to decipher the regulatory structures of the gene and metabolic networks that control the various aspects of the phenotype, including diseases. It is clear that not all biological information is contained in the DNA sequences but there are relational dynamics that are expressed at different levels of cellular organization and which are not yet known. The bioinformatics approach consists in the development of mathematical models and algorithms that allow to

extract relevant information from empirical data and to derive meaningful predictions to be subjected to experimental control [137].

2.3 Iris recognition

One of the most used physical trait in biometrics is the iris, the annular portion of the eye with a variable colour delimited by the sclera and crossed by the pupil. It is a thin membrane, visible from the front through the transparency of the cornea. The eye structure contains blood vessels, pigmented cells and two layers of smooth muscle. The contractions of these muscles allow the variation of the pupil diameter. The iris acts as a muscular diaphragm, regulating the amount of light reaching the retina; in fact, it is formed by elastic connective tissues that begin to form since the eighth month of gestation. The pigmentation and arrangement of the radial iris fibers are unique characteristics for each individual.

2.3.1 Daugman's algorithm

During the detection phase, the main problem is how to recognize the margins of a pupil inside an image. This step is crucial because it ensures that, every time, the same coordinates are assigned to certain parts of the iris in different images. The basic idea is that the pupil appears darker than the iris, with a drastic brightness change between iris and pupil. Assuming that the pupil has a circular shape, a series of concentric circumferences is traced. By adding up the brightness that each circumference has in turn for each point, the maximum value is reached when matching the centre and radius of the pupil. In this way it is possible to find the iris's edge, delimited by the sclera. In this phase, the algorithm involves the use of a integro-differential operator specifically engineered to locate the circular regions of iris and pupil. The iris is located using the following integro-differential operator:

$$max_{(r,x_0,y_0)} \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint_{r,x_0,y_0} \frac{I(x,y)}{2\pi r} ds \right|$$

$$G_\sigma(r) = \left(\frac{1}{\sqrt{2\pi\sigma}} \right) exp \left[-\frac{(r-r_0)^2}{2\sigma^2} \right]$$

where $I(x,y)$ is the image of the eye, $G_\sigma(r)$ is a Gaussian Smoothing function with center r_0 and standard deviation σ, the symbol $*$ indicates the convolution, ds indicates each arc of infinitesimal length and the division by $2\pi r$ serves to normalize the integral. The operator looks for a circular path where the variation of the pixel value is maximized, varying the radius r and the center (x_0, y_0) of the circular contour. The algorithm is particularly efficient when the iris is visible and no part of it

is covered by the eyelid, otherwise it could mistakenly identify the edge of the iris. To overcome this issue, a procedure similar to that used for the iris is adopted to localize the upper and lower eyelid line. The integration path used in the operator definition is, in this case, an arc parabolic and not a circular one. It can be described by a spline whose parameters are estimated using common statistical approaches. The splines are used to approximate the complex edges and their determination in a digital image is based on *(i)* the size of the module to be defined, *(ii)* the luminance, *(iii)* and the colour of the pixel [72]. The algorithm maps each iris point to a point with polar coordinates (r, θ), where $r\epsilon[0,1]$ e $\theta\epsilon[0,2\pi)$, and it suits pupil dilation and variability in dimension producing an invariant representation with respect to size and translation in the polar coordinate system.

During the recognition phase, the rotation is considered by translating the iris template in the direction θ until the two irises are aligned to be compared. The transformation from cartesian (x, y) to polar coordinates (r, θ) occurs as follows:

$$I\left(x\left(r,\theta\right), y\left(r,\theta\right)\right) \rightarrow I\left(r,\theta\right)$$

$$x\left(r,\theta\right) = \left(1-r\right) x_p\left(\theta\right) + r x_s\left(\theta\right)$$

$$y\left(r,\theta\right) = \left(1-r\right) y_p\left(\theta\right) + r y_s\left(\theta\right)$$

where $(x(r,\theta), y(r,\theta))$ are defined as a linear combination of the coordinates of the pupil edge $\left(x_p\left(\theta\right), y_p\left(\theta\right)\right)$, and the coordinates of the outer iris edge $\left(x_s\left(\theta\right), y_s\left(\theta\right)\right)$. The pupil and iris edges are extracted by applying the $Gaborfilter$ to the image $I\left(\rho,\theta\right)$ in the polar coordinate system.

$$G\left(r,\theta\right) = e^{i\omega(\theta-\theta_0)} e^{-\frac{(r-r_0)}{\alpha^2}} e^{-\frac{i(\theta-\theta 0)}{\beta^2}}$$

where (r,θ) indicates the position, α and β represent the dimensions of filter, and ω its frequency. Each bit h in an iris code can be regarded as a coordinate of one of the four vertices of a logical unit square in the complex plane:

$$h_{\{R_e,I_m\}} = sgn_{\{R_e,I_m\}\rho\phi} \int\int I\left(\rho,\phi\right) e^{-i\omega(\theta_0-\phi)} e^{-\frac{(r_0-\rho)^2}{\alpha^2}} e^{-\frac{(\theta_0-\phi)^2}{\beta^2}} \rho d\rho d\phi$$

where $h_{\{R_e,I_m\}}$ is a bit with complex values in which the real part and the imaginary part assume a value of 0 or 1 depending on the sign (sgn) of the 2D integral. For each coordinate element (r_0,θ_0) is calculated a bit pair (h_{Re}, h_{Im}). Instead, the values of the parameters r_0, θ_0, α, β and ω are discretized to obtain a code of 256 bytes. So a total of 2048 bits is extracted from the whole iris image (256 bytes long).

The comparison between two iris codes is based on the $HammingDistance$ between the two 256-byte codes:

$$HD = \frac{1}{N} \sum_{N}^{j=1} A_j \, (XOR) \, B_j$$

where $N = 2048 \, (256 \times 8)$ if there are no occluded areas of the iris. Otherwise, only the regions valid of the iris are considered in the calculation. Given two iris codes X and Y, their Hamming Distance is given by the sum of the discordant bits (XOR) divided by the total number N of bits in the code. If the two codes are generated by the same iris, the Hamming Distance will be close to 0, due to the high correlation. To manage possible rotations, the templates are shifted to the left and the right (bit by bit) and therefore more Hamming Distances are calculated accordingly. For recognition, the lower Hamming Distance is selected as it corresponds to the best match between the two templates.

2.3.2 Canny's algorithm

The Canny algorithm is an edge detector with a low error rate, which means it can capture most of the edges with high precision by recognizing as many real edges as possible within the image. It is mainly used for iris detection and is composed by four main stages:

1. *Noise reduction.* A primary problem in the contour recognition algorithms is the presence of noise in the unprocessed images. So as a first step a spatial filter is applied to the image. The aim is to remove the high frequencies where the noise interference is more problematic. A *Gaussian filter* is chosen because its standard deviation value can be modified to favor the recognition of small and sharp edges or larger and gradual edges. The result of this first phase is a slightly blurred image, in which no pixel is significantly affected by noise.

2. *Search for the gradient of image brightness.* Two gradients are calculated on the x and y axis by using the *Sobel operator* that returns the value of the first derivative of both the horizontal and vertical direction. This operator guarantees a fair approximation while keeping the calculation request low. Then, for each pixel the direction angle θ between gradients is computed and rounded to the values of 0, 45, 90, and 135, representing the vertical, horizontal, and the two diagonal directions, respectively.

3. *Elimination of non-maxima.* In this phase, the values of the pixels not considered part of the outline are reset, i.e., the pixels whose intensity value is not greater than that of the adjacent pixels located along the direction given by

the value θ at that point. The result is a binary image with a thin line at the edges of the objects in the image. The algorithm compares the intensity of the current pixel border with the intensity of the pixel border in the positive and negative gradient directions. If the intensity of the edge of the current pixel is greater than the other pixels of the mask in the same direction, the value will be preserved, otherwise, the value will be suppressed.

4. *Identification of the contours by thresholding with hysteresis.* A weakness concerning the contour recognition algorithms is given by the threshold value selected to obtain a good result. This value is often determined empirically, by predicting what is the minimum value that the gradient must-have for the considered pixel to be part of the outline. To mitigate this weakness, the thresholding with hysteresis method is used. This method is based on the high threshold value and the low threshold value. Each point whose gradient is higher than the high threshold is automatically defined part of the contour (value 1). Furthermore, each point contiguous to a point of the contour that has a gradient value higher than the low threshold joins the contour. Finally, a value of 0 is assigned to all the remaining points and those below the low threshold. The result is the binary image of the outlines.

Chapter 3

RELATED WORK

In this section I discuss the works related to the studies presented in this thesis, that is, works related to the support offered by Artificial Intelligence to both degenerative and oncological disease identification.

3.1 AI support to degenerative diseases

For the first group we examined Parkinson and Melanoma disease. In the case of Parkinson desease, we concentrated our attention to the application of technique based on text mining and information retrieval to patient record data. In the case of melanoma we examined previous research aiming at supporting Melanoma detection using Artificial Intelligence.

3.1.1 Related on Text Mining e Information Retrieval Tecnique

The choice to adopt Information Retrieval (IR) techniques in the medical field is increasingly common and could be a common practice in the future to support medical diagnosis.Recently, these techniques have been applied to biomedical data [45, 90, 140, 149]. In particular, Gefen and Miller [90] use the LSA on medical records related to congestive heart failure to identify patterns of associations between terms of interest. Similarly, Li and Wu [140] propose the KIP software tool for the identification of topical concepts from medical documents. An advantage of these studies is that the knowledge of the diagnosis and the treatment to be applied could be kept up to date, while one of the disadvantages could be that medical documents are analyzed without considering the real meaning of the words and this could cause errors and confusion. Mao and Chu [149] propose a phrase-based vector space model for indexing medical documents, whilst Chou and Chang [45] have developed an

IR system for clinicians and patients to retrieve similar medical case records or related documents from various databases, deriving the similarity between the concepts using their relationship based on knowledge and the similarity between two sentences was measured using their root overlaps and the similarity between the concepts. Hao et al. [100]focus on text mining in medical research. As reported, text mining reveals previously unknown new information by using a computer to automatically extract information from various text resources. Text mining methods can be thought of as an extension of data mining to text data. Word mining is playing an increasingly important role in medical information processing.

While as regards Machine Learning (ML) techniques, documents can be classified in three ways: unsupervised, supervised and semi-supervised methods, these techniques are widely used for the extraction of knowledge in biomedical data [8, 22, 24, 39, 122, 248]. In particular Beam et al. [18] presents a new reference methodology based on statistical power specifically designed to test incorporations of medical concepts, called cui2vec. This study, however, has limitations as most of the sources of health data are not easily shared, which limits the study to small local data sources. Finally, they provide a downloadable set of pre-trained embeds as well as an online tool for interactive exploration of the embeds. Dynomant et al. [70] used the Doc2Vec algorithm to train models that allow you to vectorize documents on the PubMed database to analyze whether you can replace the statistical model PubMed Related Articles (pmra). This algorithm was able to link documents sharing MeSH labels in a similar way the pmra did. Chen and Sokolova [41] they used Word2Vec and Doc2Vec unsupervised to analyze sentiment summary reports. They aimed to detect whether there is any latent prejudice towards or against a particular disease. They used SentiWordNet to establish a golden sentiment standard for data sets and evaluate the performance of the Word2Vec and Doc2Vec methods. The visualization of the data is very important, especially in the medical field, it is used to synthesize all the information relating to a patient but in particular, the visualization of the data is used to support decisions and diagnosis [193, 233]. Especially Lesselroth et al. [138] underline some problems concerning the management of information at the point of care and propose strategies for a better visualization of data including multimedia displays, clinical dashboards, concept-oriented views, metaphor graphics and probability analysis. Ropinski et al. [196] examine glyph-based visualization techniques that have been exploited when viewing spatial multivariate medical data. To classify these techniques, they derive a taxonomy of the properties of glyphs that is based on classification concepts established in the information display. Blaas et al. [21] presented to highly interactive, coordinated view-based visualization approach that has been developed for dealing with multi-field medical data. This type of visualization is based on intuitive interaction techniques and integrates analysis techniques from pattern classification to guide the exploration process. Similarly, the studies by dos Santos et al. [69] focus on applying data mining and machine learning (ML)

techniques to public health problems. As stated in this research, public health can be defined as the art and science of preventing disease, promoting health and extending life. Using data mining and ML techniques, it is possible to discover new information that would otherwise be hidden

3.1.2 Related on Iris ricognition in patological state and descrimination in IoBT service

Last decade has witnessed a growth in the use of biometric techniques to identify people. Unlike traditional authentication methods, based on information manually provided by users, e.g., password, PIN, and ID cards, *biometric recognition* relies on physiological or behavioral identifiers, that are much more difficult to be lost, stolen, copied or falsified. Thanks to these features and the rate and breadth of technological developments, identification systems [60] exploiting physical traits, e.g., face, iris, and fingerprints, are rapidly spreading[1]. Despite their indisputable strengths, biometrics raises practical, ethical and legal issues. Most debated concerns involve personal data protection, privacy and the risk of covert surveillance [33]. The discriminatory potential of technology has been for a long time a key topic in the debate on the social impact of technology. As highlighted in dating back works [235], human artifacts do not only contribute to increase efficiency and productivity in the society, but often incorporate power, authority and values that could produce discrimination [236]. Specifically, biometrics systems can only deal with people who fall within the range defined as "normal" by the system's commissioners, designers and administrators [224]. However, for a variety of reasons (e.g., injuries, age, disability, surgeries or genetic defects), people can lose or be lacking the physiological characteristics that, according to the biometrics perspective [114], must be instead universal, unique, permanent and collectable. Fingerprint recognition, for example, can be altered by specific medical conditions (chemotherapy often induces acralerythema that can delete fingerprint [5]) or age (elder people tend to have drier skin), and as consequence, they generate more false rejections than average when trying to authenticate. Similarly, voice recognition can be fooled by cold or laryingitis, while a simple swelling can cause false rejections in face or hand palm recognition [160].

Implications, in all these cases, are far reaching [225]. When biometric systems fail to cope with human features falling outside a specific range, they categorize and exclude people with consequences that become the more significant, the larger the number of contexts involved in the use of these systems. Another context to consider is the one resulting from the models of legal regulation that are emerging in the information society. Whereas social, economic and personal interactions are mediated by ICT, technology goes from being simply a subject of regulation to being a "regulatory tool". According to the paradigm of techno-regulation, the "*inten-*

[1] https://bit.ly/2RUTrkQ

tional influencing of individuals' behavior by building norms into technological devices" [136], legal safeguards are ever more going to be implemented and enforced through hardware and software tools. Highly debated in recent scientific literature, techno-regulation has given rise to a variety of applications [49, 121, 139]. In such a scenario, the certain identification of individuals deserving protection or responsible for illegal conducts is a *sine qua non* to enforce norms, and link legal consequences to given facts.

Observe that, in all the cases considered, what is looming is the risk of an exclusion from basic services and important legal safeguards with potential spillovers on fundamental rights. The answers must be seek into the creation of biometric tools diversity-aware capable of "embedding" in the technology the safeguards against exclusion. Few works which try to overcome biometrics limitations have been proposed in the literature. There have been efforts made for the biometric authentication of elderly [129], which is a new trend in biometrics. Furthermore, in [89] authors described a method exploiting CNN to detect contact lenses which can be also used to contrast spoofing attacks, while in [173] an eye-tracking method for people with motor disabilities is proposed.

Instead, when Biometric recognition is adopted for improving the security in the Internet of Things (IoT) field, it is commonly named Internet of Biometric Things (IoBT). However, despite its advantages these specific physical conditions would lead to not satisfy the requirements commonly used for biometric recognition. As a consequence, the limitations of current biometric systems can exclude a person from the use of IoBT services. Nowadays, the number of Internet of Things (IoT) services and devices is growing exponentially. However, security remains the major concern for these services. Indeed, the authentication of users is required when they connect to IoT devices. To overcame this issue, new approaches and services based on the use of biometric authentication have been proposed [128, 205], which will greatly extend the range and target of biometric applications by exploiting the recognition of physiological or behavioral characteristics (e.g., facial, iris, and fingerprint) available on the latest technological devices [119].

Nowadays, the number of Internet of Things (IoT) services and devices is growing exponentially. However, security remains the major concern for these services. Indeed, the authentication of users is required when they connect to IoT devices. To overcame this issue, new approaches and services based on the use of biometric authentication have been proposed [128, 205], which will greatly extend the range and target of biometric applications by exploiting the recognition of physiological or behavioral characteristics (e.g., facial, iris, and fingerprint) available on the latest technological devices [119].

Biometrics systems can only deal with people who fall within the range defined as "normal" by the system's designers [224]. However, for a variety of reasons (injuries, age, disability, surgeries or genetic defects), people can lose or be lacking

the physiological characteristics that, according to the biometrics perspective [114], must be instead universal, unique, permanent and collectable. This represents a concern when entire categories of individuals [30] may incur in the alteration of their physical traits increasing the risk of being excluded from biometric recognition processes.

A first study on the effect of ocular pathology on the iris recognition and on the failure due to eye diseases in iris recognition systems, was proposed in [11]. Results revealed that iris recognition performance was remarkably resilient to most ophthalmic disease states, including corneal edema, iridotomy and conjunctivitis.

In IoBT systems, iris recognition is largely adopted to make IoT more secure, often together with other biometric measures, such as ECG [119]. ecause of its uniqueness features extracted from the image of the iris are stored in a database and used to build effective and precise biometrics systems for the identification of individuals, based on information taken from databases containing characteristics extracted from the images of the iris.

Iris is a stable structure throughout the life of the individual and is normally not subject to accidents; even a surgical operation cannot alter biometric recognition as long as an intact surface of about 60% of the iris is preserved. Biometric technologies that make use of the iris are a valid and effective alternative to the development of application-oriented access control procedures, especially in the category of non-invasive identification methods [113].

Many approaches for iris detection based on Neural Network exist in the literature. For example [247] and [156], reaching accuracy of 98,43% and 95,5%, respectively. [231] proposed a deep-learning-based report generation model for the classification and screening of ophthalmic disease of fundus images. They obtained an accuracy of 96.87%. [204] proposed an overview of the applications of deep learning for ophthalmic diagnosis using retinal fundus images. A review of the application of Machine Learning techniques to ophthalmic diseases is reported in [217]. Daugman's algorithm actually remains the most accurate algorithm, largely used for this type of recognition [37, 94, 176, 206] together with the Canny edge detection algorithm [97, 167].

3.2 AI support to oncological diseases

In the case of oncological disease, we discuss related work concerning the detection of Melanoma and Leukemia.

3.2.1 Melanoma detection

The automatic recognition of melanoma is one of the most widely discussed topic in the last thirty years. Since 2012, deep learning has been applied to classification,

segmentation, localization of melanoma.

Several works adopt artificial intelligence for classifying skin lesions.

Philiips *et al.* [178] evaluated the use of an ad-hoc developed neural network named Deep Ensemble for Recognition of Melanoma to detect malignant melanoma from dermoscopic images of pigmented skin lesions. The diagnostic accuracy achieved a ROC area under the curve (AUC) of 0.93 (95% confidence interval: 0.92-0.94), and sensitivity and specificity of 85.0% and 85.3%, respectively.

Ercal *et al.* [73] present the detection of melanoma using a neural network for the automatic separation of melanoma classifying three benign categories of tumors that have characteristics similar to melanoma. The approach uses discriminating characteristics, based on the shape and colour of the tumor. Nasr-Esfahani *et al.* [164] propose a method where clinical input images are pre-processed as they may contain artifacts or disturbing elements. Subsequently, the pre-processed images were given as input to a pre-trained CNN. The CNN classifier, which is trained by a large number of training samples, distinguishes between cases of melanoma and benign cases. Namozov *et al.* [163] propose a deep neural network model with adaptive linear units. The experimental results show that a CNN model with linear units with parameterized adaptive sections exceeds the same network with different activation functions in the melanoma classification task. Yu *et al.* [243] propose a new method for the recognition of melanoma by exploiting very deep CNN. Therefore, they create a completely convolutional residual network (FCRN) for accurate segmentation of the skin lesions and further improve the network classification capacity by incorporating a framework for integrating contextual information on multiple scales. Finally, they integrate the FCRN (for segmentation) and other very deep residual networks (for classification) to form a two-stage framework. Majumder *et al.* [148] propose a method to extract the geometric characteristics distinct from the dermoscopic images to classify benign and malignant melanomas. Variation between melanoma and non-melanoma images is computed by extracting the difference between maximum and minimum Feret diameters of the best fit ellipse to skin lesion. Brinker *et al.* [28] propose the use of a CNN to compare the performance of this trained network exclusively with clinical dermoscopic images with the manual classification of dermatologists.

Sultana *et al.* [214] present a review about the deep learning techniques to detect melanoma cases from the rest skin lesion in clinical and dermoscopy images. However, the selected techniques have been applied on different datasets with different pre-processing. The results are shown in Table 3.1 taking in to account the adopted neural network techniques and the selected dataset. On the contrary, our proposal performs a comparison of neural network techniques by choosing one dataset on which we applied a specific pre-processing phase.

Several research efforts have been devoted to provide support in the melanoma diagnosis by identifying relevant features and setting the alert thresholds [9, 52].

Table 3.1: Comparison with different neural networks on different dataset.

Methods	Dataset	Accuracy	Sensitivity	Specificity
Fine-tuning using VGGNet	ISBI 2016	0,81	0,79	-
CNN + ResNet	ISBI 2016	0,85	0,54	0,94
CNN + Fisher-encoding	ISBI 2016	0,83	-	-
Multi-resolution CNN	Dermofit	0,79	-	-
Multi-scale feature extraction CNN	Dermofit	0,82	-	-
CNN (5 layer)	MED-NODE	0,81	0,81	0,80
ResNet + Bilinear-pooling	MoleMap	0,71	-	-
ResNet + Bilinear-pooling	ISBI 2016	0,85	-	-
Deep learning + Sparse-coding + SVM	ISIC	0,74	-	-
Deep learning ensemble	ISIC	0,81	0,69	0,84

In [109] a mobile system performing image recognition and classification has been proposed.

Kassianos *et al.* [120] identified 40 smartphone apps to detect or prevent melanoma by nonspecialist users. None of them was based on AR.

Barata *et al.* [15] examined two systems for melanoma classification. The first system uses global methods to classify skin lesions (ABCD rules), whereas the second system uses local features, such as gradient histograms and color histograms of reduced regions. Global methods reached sensitivity of 96% and specificity of 80%, while local methods obtained sensitivity of 100% and specificity of 75%. The dataset was composed by 176 dermoscopy images.

Abuzaghleh *et al.* [3] proposed a system for supporting the prevention of skin burn due to sunlight and an image analysis component that classifies the image using the PH2 dataset. The features extracted were 2-D Fast Fourier Transform, 2-D Discrete Cosine Transform, Complexity Feature Set, Color Feature Set and Pigment Network Feature Set. The experiment results revealed that the classification reached 96.3%, 95.7%, and 97.5% of accuracy for benign, atypical, and melanoma skin lesions, respectively.

In [35], a mobile app is proposed to assist melanoma detection by using a CNN. The dataset was composed of smartphone lesion images and lesion clinical information. Due to the rarity of dermoscopic images, an evolutionary algorithm has been adopted to balance datasets. A balanced accuracy of 92% has been reached.

Pacheco *et al.* [171] created a dataset with smartphone clinical images and patient clinical data. This information was proposed as input to deep learning models, such as CNN, to combine features from images and clinical data. Then the model performance with and without using clinical data has been compared. The accuracy was improved of 7% when additional information is considered.

Hoang *et al.* [68] adopted a dataset of images taken by a mobile device, but the

dataset is not accessible. It is devoted to self-diagnostic and the end-users are the patients. The classifier exploits numerical features to characterize a skin lesion which are a reduced set of features obtained by feature selection.

We also explored the app in the stores. In particular, DermEngine [64] monitors the nevus evolution: it enables the clinician to associate a nevus picture to a specific position of the human body, in such a way as to retrieve the nevus in the future. An additional camera has to be added to the mobile device.

SkinVision [147, 177] is a smartphone app that takes the lesion's picture by the phone's camera and provides the risk assessment of the lesion. The risk analysis algorithm of SkinVision is based on gray-scale image analysis and the fractal dimension of skin lesions. Recent data on a sample of proprietary smartphone images show that it reaches sensitivity of 97% and specificity of 78%.

LÅ«bax [42] exploits a proprietary database composed of 12,000 images of lesions certified by dermatologists. The approach creates a single high-dimensional signature for each image on the base of lesion features (i.e., size, color and shape) and a computer algorithm which compares the characteristics of new images with images in the database to identify the nearest-match diagnosis. It got sensitivity of 90.4%, specificity of 91.5%, and accuracy of 90.8%.

3.2.2 AI for feature selection of deferentially expressed genes

Neural networks are powerful machine learning methods that are often used to learn data representations at multiple levels of abstraction. These representations are useful for many applications such as reconstruction, classification, grouping, and recognition. Prediction models use the neural network capabilities to classify, group samples, or apply statistical analysis [53]. In particular, neural networks are also commonly used to build cancer prediction models from microarray data [53]. The high dimensionality of gene expression profiles is a crucial problem in building these models. To minimize the feature size and maximize the classification performance a feature selection pre-processing phase has to be adopted. It is a type of multi-objective optimization problem.

Feature selection on microarray data is an area currently very explored to discriminate a subset of optimal features of the various existing classifiers to obtain maximum accuracy. In the following, we discuss the main results of feature selection approaches applied on microarray data, summarized in Table 3.2, where we report for each approach the considered datasets and classifiers, and the average accuracy.

Chen *et al.* [40] adopted a Kernel-based clustering methods (KBCGS) for gene selection. They compared the performance of their approach with other algorithms to select an excellent number of features. The Maximum-Minimum Cross-Entropy Criterion [158] is used to determine the best method.

Recently, different approaches have been developed to perform gene selection

on a genetic dataset.

Dhrif *et al.* [65] presented a new variant of the Particle Swarm Optimization (PSO) algorithm to increase the classification accuracy and preserve the acceptable dimensions of feature subsets when there are many uninformative data. For this purpose, a new encoding scheme is used for mapping particle positions to probabilities. The aim is to expand the search of features in a continuous space without limiting solutions to local optima. To test the stability and scalability of the algorithm they created synthetic datasets.

Kang *et al.* [118] proposed a relaxed Lasso-Gen (rL-Gen) method for tumor classification in which the dataset is first z-scored normalized, then a relaxed Lasso is applied for gene selection and, finally, a generalized multi-class support vector machine (Gen) is used as a classifier.

Ghosh *et al.* [92] proposed a recursive meta-heuristic model is called Recursive Memetic Algorithm (RMA), inspired by Dawkin's notion of meme. The proposed Recursive Memetic Algorithm (RMA) model improves classification accuracy and has a higher convergence rate in finding the cancer biomarker compared to other meta-heuristics such as genetic algorithm (GA) or basic MA.

Sun *et al.* [215] presented a global feature selection method based on a semidefinite programming model relaxed from the quadratic programming model with maximization of feature relevance and minimization of feature redundancy, i.e., Minimum Redundancy Maximum Relevance (MRMR).

Table 3.2: Features selection approaches applied on microarray data.

Dataset	Approach	Classifier	Average accuracy
ALL, AML, DLBCL, Lung, Prostate, Lymphoma, SRBCT, Brain, NCI60	KBCGS [40]	SVM, KNN	93,45%
Leukemia, Prostate, B-cell Lymphoma	PSO [65]	Random Forest	97,22%
DLBCL, CNS, Lung, Ovarian, Brain, Lymphoma, MLL, TOX171	Relaxed Lasso [118]	rL-GenSVM, KNN	96,43%
MLGSE2191, Colon, DLBCL, Leukemia, Prostate, MLL, SRBCT	RMA GA [92]	SVM	95,86%
AML, ALL, Breast, Colon, DLBCL, Lung, Medulloblastoma, Prostate	MRMR [215]	CART, Naive Bayes, Random Forest	83,41%
SRBCT	Gene Masking [197]	Nearest Shrunken Centroid Classifier, Nearest Centroid Classifier	100%
Leukemia, Colon, DLBCL, Prostate, Wang Breast, Lung Adenocarcinoma, Medulloblastoma	Multi-objective heuristic algorithm [144]	SVM	87,71%
Ovary, Lung, SRBCT, CNS, DLBCL, Prostate, Leukemia	Evolutionary Operators [170]	Wilcoxon	73,60%
Colon, Prostate, Lung	PCA [32]	ANN, GAHI	86,33%
Leukemia, Colon, Lung, Ovarian	Hybrid GA + PSO [237]	ANN	98,63%
Colon, Adenocarcinoma, SRBRT, NCI60	DPCAForest [63]	SVM, Recursive Feature Elimination	90,25%

Saini *et al.* [197] proposed a gene masking derived from the genetic algorithm. An optimal gene mask is searched that provides the largest performance gain by removing the largest number of features for the chosen classification algorithm.

Lv *et al.* [144] applied a multi-objective model following the analytic hierarchy process that gives more importance to the detection accuracy than the feature size to build a model such as the multi-objective optimization algorithm (MOEDA). This solution is based on a type of distribution estimation algorithm (EDA) that guides

the search for the optimum by building and sampling explicit probabilistic models of promising candidate solutions.

Othman *et al.* [170] proposed and developed multi-objective hybrid cuckoo research with evolutionary operators for gene selection. The evolutionary operators used are double mutation and simple crossover operators. The results of the experiments revealed that the developed algorithm, multi-objective cuckoo search with evolutionary operators, outperformed the cuckoo and multi-objective search algorithms with less significant selected genes.

Cahyaningrum *et al.* [32] proposed a technique based on Principal Components Analysis (PCA) to select the most relevant features. Moreover, they proposed the use of ANN and e GA Hybrid Intelligence (GAHI) for cancer detection. Although ANN is recognized as one of the methods to classify microarray data, GA is used in this case to optimize the ANN architecture.

Wu *et al.* [237] proposed an ANN classifier. To initialize the structure, an algorithm was used to choose input variables on layered links and different activation functions for different nodes. Then, a hybrid method integrating GA and particle swarm optimization (PSO) algorithms were used to identify an optimal structure with the parameters encoded in the classifier.

Deng *et al.* [63] proposed DPCAForest, a deep forest-based model that integrates the deep forest and the component analysis of the dynamic principle. DPCAForest adaptively generates minority samples based on sample distribution and then performs principal component analysis dynamically synchronized with the growth of the deep forest to reveal the important features with the highest variance. With dynamic PCA, the model can perform feature extraction in a data-driven manner based on cross-validation and obtain information on the merging between layers.

Various lines of research use the evolutionary calculus to develop solutions to selection problems. Recently, metaheuristic algorithms have been used to perform genetic selection and their implementation has been studied. However, despite the various methods proposed for genetic selection, they suffer from local and optimal stagnation problems and high computational costs, which therefore cannot guarantee the optimal and reasonable use of metaheuristic algorithms in a wide range of research of identified genes [112].

In our approach, unlike others we use a Bayesian inference method to perform the feature selection on the dataset and autoencoders to perform a second feature selection applied on the genes differentially expressed.

Chapter 4

ARTIFICIAL INTELLIGENCE IN DEGENERATIVE DISEASES

The aim of this chapter is to answer the following research question:

RQ1: how Artificial Intelligence may support the degenerative disease identification?

Specifically, I will deal with Parkinson's disease and Coloboma of the eye. In the former, in (C01) we used IR and ML techniques to identify eventual correlations between documents to recognize the different classes of patients based on the specific medical reports. Furthermore, we perform semantic analysis, Text2Vec, and Doc2vec techniques on the medical report to highlight the most characterizing keywords for Parkinson's disease (J01). Finally, a visualization system was adopted to support diagnosis for clinicians. While regarding the Coloboma of the eye, we focused on the difficulty of iris recognition when it is affected by Coloboma, a congenital abnormality of membranes of the eye. We show how this pathological state impacts the quality of the result of the Daugman and Canny edge detection algorithms, which represent the most widespread methods used for the iris localization step in eye-based biometric (C02- J02). Moreover, I also propose an ongoing research that extends the Daugman algorithm to people affected by Coloboma.

4.1 Supporting Parkinson's detection

In this section I discuss the results of my thesis related to the Parkinson's disease. First I summarize main conceptsrelated to this disease and the adopted dataset, then I present the proposed process to identify correlations among Biomedical Data through Information Retrieval and Machine Learning Techniques.

4.1.1 Parkinson's disease

(PD)is a neurodegenerative disease, belonging to the "*Movement Disorders*" category. It is originates from the degeneration of neurons in the brain that produce the neurotransmitter "$dopamine$". In the early stages of the disease, the most obvious symptoms are related to movement, and include tremors, stiffness, bradykinesia, postural instability, slowness in movement and difficulty walking. Afterwards, cognitive and behavioral problems may arise, such as dementia, depression, psychotic features, autonomic dysfunction, oculomotor abnormalities [91]. In particular, in PD the production of dopamine in the brain decreases consistently and the reduced levels of dopamine are due to the degeneration of neurons in an area called *substantia nigra* (cell loss is over 60% at the onset of symptoms). Moreover, from the marrow to the brain accumulations of a protein called alpha-synuclein begin to appear. This insoluble protein accumulates within neurons forming inclusions, called Lewy bodies [59]. The causes of PD are not yet known but, it seems that there are multiple elements that contribute to its development. These factors are mainly:

- Genetic mutations: among these, the mutation of the $LRRK\ 2$, named also $PARK8$, is the most relevant. The heterozygous mutation, $2877510\,G \rightarrow A$, of this gene is the most commonly described, representing the majority of familial cases of cases of idiopathic PD [59].

- Toxic factors and work exposure such as some insecticides or herbicides.

The diagnosis of PD remains a clinical diagnosis because there are neither objective tests nor specific biochemical and neuroradiological markers. However, in the last decade one of the objectives of the research has been to improve the specificity of the classical diagnostic criteria.

The diagnosis of Parkinson's disease remains a clinical diagnosis since there is no objective test or specific biochemical and neuroradiological markers. In the last decade, however, one of the research objectives has been to improve the specificity of classical diagnostic criteria: in fact, the "United Kingdom Parkinson's disease Society Brain Bank" has proposed clinical criteria that are still widely used in clinical practice and research protocols. These diagnostic criteria establish that the sign necessary to diagnose Parkinson's disease is bradykinesia or akinesia, associated with at least one of the other so-called major signs mentioned above, i.e. muscle stiffness, tremor at rest and postural instability. These diagnostic criteria underline how clinical diagnosis is based on the combination of some "cardinal" motor signs and on the exclusion of symptoms considered "atypical" [91]. In conclusion, the symptoms of Parkinson's disease manifest themselves differently in different patients, who may experience some symptoms and not others, and also the rate at which the disease progresses varies from individual to individual. For this, the misdiagnosis rate can be relatively high.

4.1.2 Dataset

The dataset used is the Parkinson's Progression Markers Initiative (PPMI) dataset, it is the result a clinical study based solely on observations aiming at fully evaluating significant cohorts of interest by using advanced imaging, biological sampling, clinical and behavioral assessments to identify biomarkers related to the progression of PD. The collected data may be helpful in the research of therapies to slow down or stop this progression. The activity conducted by PPMI is an "open source" study, the data and samples collected and acquired by volunteer participants, affected and not by the disease, will allow the development of a database and a complete biorepository, which is currently available online and updated. every eight months. Being data collected from patients from various continents one of the main tasks of PPMI is to coordinate the management of the various data, defining a protocol for the collection and coding of data. The elaborated repository can be downloaded by accessing the portal of the PPMI site to allow the scientific community to conduct complete and exhaustive research.

The dataset consists of files in CSV format, containing information about six macro-areas listed in the following:

- *Biospecimen*: collection of data related to clinical tests, such as blood collection, DNA and lobar puncture.
- *Imaging*: use of imaging techniques, such as Magnetic Resonance, Pet and DatScan through which it is possible to observe non-visible areas of the organism.
- *Medical History*: clinical history of patients from the first symptoms of the disease to the latest health conditions. The collection includes possible side effects of the medicines taken, results of neurological examinations, physical and so on.
- *Motor MDS-UPDRS*: collection of motor disturbance data through the use of the MDS-UPDRS scale to evaluate the stage of Parkinson's disease.
- *Non Motor Assessments*: collection of data related to cognitive and emotional-behavioral disorders.
- *Study Enrollment*: collection of conclusive data on particular studies conducted on patients.
- *Found*: collection on personal habits and lifestyles data.

Figure 4.1 shows the data model of the PPMI dataset, in which there are five entities:

- *Patient*: represents the set of patients participating in the study.
- *Event*: represents the set of tables that refer to the visits and analyzes to which patients are subjected.
- *Biospecimen Analysis Result*: represents the set of tables in which the analysis of the results for the controls to which the patients have undergone are present.
- *Family History*: a set of tables that describe the patient's family histories.
- *Medication*: a set of tables in which the medicines taken by patients are cataloged.

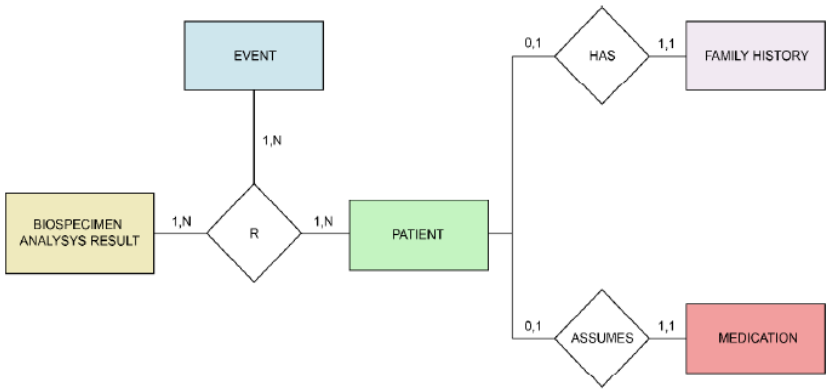


Figure 4.1: The data model of the PPMI dataset.

The entities described above are connected through relationships:

- *R*: represents the relationship that exists between the entities, $Event$, $Patient$, and $Biospecimen\ Analysis\ Result$, through the $PATH$, that is the unique attribute that identifics the patients;
- *HAS*: is the relationship that represents the connection between patients and their family history;
- *ASSUMES*: is a relationship that associates to each patient the medicines he takes.

4.1.3 The proposed process

We present the process that allows us to find the correlation between the information on the visits and the patient's disease status, which may be: sick (PD and GENPD), healthy (HC, GENUN and SWEDD) and healthy with typical symptoms of the disease (PRODOMAL, i.e., subjects suffering from insomnia and have mutations of the $LRRK\ 2$ gene). The version of the PPMI data used in the first paper is updated in July 2018 and the complete dataset consists of 113 files in CSV format [174]. The process is articulated in the following phases:

Step 1 After an initial analysis of the dataset, a skimming of the tables was carried out by selecting only those of interest for the text analysis. This selection was made only on the tables in which symptoms strongly correlated with the disease appear, excluding all those containing diagnostic information. Moreover, we evaluated only the $screening\ visits$ (SC), the $basic\ line\ visits$ (BL), the $visits\ 1{-}14$ (V01-V014), the $symptomatic therapy\ visits$(ST) and the $adverse\ events\ visits$ (LOG). The process examines them in chronological order.

Step 2 Following the selection and recognition phase, a modification of the tables is made, transforming in textual form only the columns related to the relevant symptoms, in which the answers of the patients to the various questionnaires or of the clinicians were present in numerical form. Subsequently, the columns that did not contain any type of relevant information regarding symptoms such as numeric and text fields containing abbreviations, which may vary from table to table have been completely eliminated. Finally, a further skimming was carried out by eliminating all the tables that gave reliable information on the diagnostic status of the disease.

Step 3 At this point the real process of correlating information begins. It was made sequentially as the visits progressed, adding the records of the next visit to previous ones. For each collection of documents taken into consideration, information is extracted for each individual patient. All the extracted information is kept in a new collection (Corpus). Subsequently, an initial cleaning of the text is performed. In particular, we performed stopword removal by also excluding words specific of the disease, such as "parkinsonian", "parkinsonism" and "parkinson", because they could be discriminatory terms for the classification. We also used a stemming algorithm to transform the words in their root form, called "theme". Then, we created the $n - by - m$ document-terms matrix A, where a generic entry $A_{i,j}$ denotes the number of times that the i_{th} term in the j_{th} document appears. For the weight associated to each pair (term, document) we used the term frequency-inverse document frequency, also known as $tf - idf$ [67]: $tf\_idf(A) = log(tf(A)) * idf(A)$. In particular, for every term t_i and document d_j in A, tf_idf is computed as follows:

$$tf\_idf(A[t_i, d_j]) = log(A[t_i, d_j] + 1) * ln\left(\frac{|d|}{\sum_{i,j} A[t_i, d_j] > 0}\right)$$

The tf_idf is a function used in information retrieval to measure how important a word or a document is in a corpus [185]. The tf-idf value proportionally increases the number of times a word or document appears in the corpus. Subsequently, we have applied LSA (Latent Semantic Analysis) to the A matrix. LSA assumes that there is a latent structure in the use of words that can be hidden by the words used in a document. To this aim, a singular value decomposition (SVD) is applied to the A matrix [101] to decompose it in the product of three matrices, TSD^T. The matrix S is an $r \times r$ diagonal matrix of singular values and T and D have orthogonal columns. SVD also provides a simple strategy for optimal approximate fit using a subset of k concepts (the space of the underlying concepts) corresponding to the largest singular values in S. The selection of a "good" value of k (i.e., the singular values of the dimensionality reduction of the concept space) is an open issue and a number of strategies have been proposed in the past (e.g., percentage of number of terms, fixed number of factors, etc.). In our approach, k is computed according to the Guttman-Kaiser criterion [98]. Once such a semantic space is created, the lexical distance among terms and among combinations of terms can be calculated by projecting them onto that space [90]. The documents are represented by vectors in the latent semantic space and the cosine between each pair of vectors indicates the similarity or dissimilarity between the documents. The value of the cosine varies from -1 (when the two documents are different from each other) to 1 (when the content of the document is the same).

$$d\,(i, j) = 1 - cos(V_i, V_j)$$

where i and j represent the documents and V_i and V_j the corresponding vectors in the latent semantic space. This distance takes values ranging from 0 (when the content of two documents is the same) to 1 (when the content of two documents is different).

Later [83] we have updated the dataset and complete dataset consists of 145 files in CSV format, then we applied:

– *Tex2Vec*: Tex2Vec [203] is technique of text analysis and Natural Language Processing (NLP) by building machine learning algorithms based on text data which main goal is to provide an efficient framework with concise APIs for text analysis. It is built around streaming APIs and iterators, which allows the construction of the corpus from iterable objects. This analysis

allows us to build a matrix of document terms (DTM) and to elaborate the text by creating a map from words in a vector space. This technique is based on the concept of Word Embedding, a methodology of natural language processing to map words or phrases present in vocabulary, in a corresponding vector of real numbers, used to discover semantic correlations between them. To identify similar documents, we use cosine similarity identically. Also, in this case, a corpus of documents is built by selecting only those that coincide with the context in which the phenomena of interest reside, since even the inclusion of a large collection of high-quality documents could fail if the context of these documents does not is aligned with the phenomenon of interest. Moreover, it is necessary to have a very large corpus to create a representative sample and to increase the chances of a word appearing in it. In this case, the stemming and the SVD are not applied. To represent documents in a vector space, we need to map terms with identifiers. In such a way as to represent a set of documents as a sparse matrix, where each row corresponds to a document and each column to a term. In our case, we have created a Document Term Matrix based on vocabulary. Doing nothing but cataloging the unique terms and assigning a unique ID. After, the similarity matrix is calculated using the DTM applying the cosine similarity. The comparison between documents with the cosine similarity also takes place. These are used to find which vectors are most similar to each other and which documents have a similarity greater than a specified threshold.

– *Doc2Vec*: Doc2Vec [135] is a technique that allows you to transform textual documents into vectorial representations that protect their semantics, trying to keep all possible information expressed in the text within the vectors, for example managing to interpret the information of similarity or thematic diversity between various text blocks. In reality, the Doc2Vec is an evolution of the Word2Vec technique, which consists of a group of models, used to do word embedding, whose purpose is to translate words or sentences into vectors of real numbers, or, in a form easily computable by compilers and which it manages to represent are not the word intended as a "sequence of characters" but also the meaning it assumes, thus managing to create a coding that allows, for example, to summarize concepts of similarity or opposition about other terms. These models are nothing more than two-level neural networks trained, through an unsupervised approach, to reconstruct the linguistic contexts of words; Word2Vec takes as input a large fragment of text and builds a vector space in which each word is uniquely assigned to a corresponding vector in space. The goal of Doc2Vec is to create a numeric representation of an entire document regardless of its length.

The purpose of Doc2Vec for its similarity to Word2Vec is to create a vectorial representation of an entire document regardless of its length, therefore, the vectors obtained will summarize the main theme or the global meaning of the entire document. It makes use of the Word2Vec model and in input, another vector is added, called DocumentID. So after training the neural network, you will have not only the word-vector (the vector representation of the words) but also a document vector (vector representation of the document). The purpose is simple, taken as input the DocumentID, the model uses the similarities between the words learned during the training (the word-vector) to build a vector that will include the words contained in it. By comparing these vectors, for example using the cosine similitude, we can then compare multiple documents with each other to verify their similarities. Doc2Vec, according to the Word2Vec approach used as a base, is divided into two methodologies; in particular, we have the "Distributed Memory version of Paragraph Vector" (PV-DM) deriving from CBOW and the "Distributed Bag Of Words memory version of Paragraph Vector" (PV-DBOW) deriving from Skip-Gram. In particular, we based ourselves on the Word2Vec Skip-Gram, where the task of the neural network is to calculate, given an input word, the probability for each word of the vocabulary (together with all the words obtained from the training documents) to be close (juxtaposed within the text) to it. In reality, the concept of "closeness" between words is described through the definition of a measure, called windows size, which describes the number of terms to be analyzed, preceding or following the word given as input.

The network, therefore, for each input word, will have to find the probability that that specific word forms a pair with another word of the vocabulary; therefore, the net will be trained according to the number of times each pair is used. To allow training of the neural network it is necessary to provide a numerical representation of the words, as these, in the form of strings, cannot be easily used; for this reason, a vocabulary of the words obtained from the training documents is built and one-hot vectors will be used as input to the neural network; vectors of size equal to the size of the vocabulary, consisting of all negative bits except a positive one in correspondence with the reciprocal term in the vocabulary. The output of the neural network will also be a vector of the same size, as it will also use the indices to refer to the terms of the vocabulary, but, it will contain the probabilities of the various terms of being "close" to the word given in input.

After running the LSA, Text2Vec or Doc2Vec algorithms and obtaining the similarity matrix between the various documents, a clustering technique can be

applied. In this work we use two types of techniques to compare categorization on the type of patients: the k-means [6] and the Fuzzy c-means clustering [88]. The main difference lies in the way in which the classification of the elements takes place:

- In the k-means technique the elements can belong only in mutual exclusion to a cluster and once assigned to a given cluster they can can no longer be moved. The k-means algorithm is part of the "hard clustering" techniques and it is a partition clustering algorithm that allows to subdivide a set of objects in K groups based on their attributes, by partitioning the data set into unique homogeneous clusters whose observations are similar but different from other clusters. The k-means iteratively improves the initial centroids by minimizing the total intracluster variance, ie maximizing the similarity between the documents. The resulting clusters remain mutually exclusive.
- In the Fuzzy c-means technique the elements can belong simultaneously to both clusters, without any constraint. Fuzzy c-means clustering, also referred to as soft clustering or soft k-means where each element has a set of membership coefficients corresponding to the degree of the link with a given cluster; this value can vary from 0 to 1. The Fuzzy c-means algorithm is one of the most common fuzzy clustering algorithms, the centroid of a cluster is calculated as a weighted average of all points, based on the degree of cluster membership. The clustering process is accomplished through an iterative optimization of the following function:

$$\sum_{v=1}^{nc} \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} u_{iv}^{r} u_{jv}^{r} d(e_i, e_j)}{2 \sum_{j=1}^{m} u_{jv}^{r}}$$

  where e_i and e_j are pairs of entities selected in the set of all cluster entities. The size of this set is m, while nc is the number of clusters to identify and u_{iv} is a not negative value that specifies the membership of the entity e_i to the cluster v. The sum of all the relevances of a given entity e_i is 1, while the exponent of membership is r and can assume values between 1 and ∞. In the case r is close to 1, the behavior of the algorithm is similar to that of the k-means algorithm. The clustering process will stop when the inequality occurs [195]:

$$max_{i,v=1\ldots nc} \left| u_{iv}^{t+1} - u_{iv}^{t} \right| < \varepsilon \tag{4.1}$$

t indicates the maximum number of iterations, while ε represents a termination criterion. The value for ε in [0,1]. Fuzzy c-means computes a membership matrix that is used to generate clusters. We empirically set $r = 1.01$, $\varepsilon = 1e^{-20}$ and $t = 1000$.

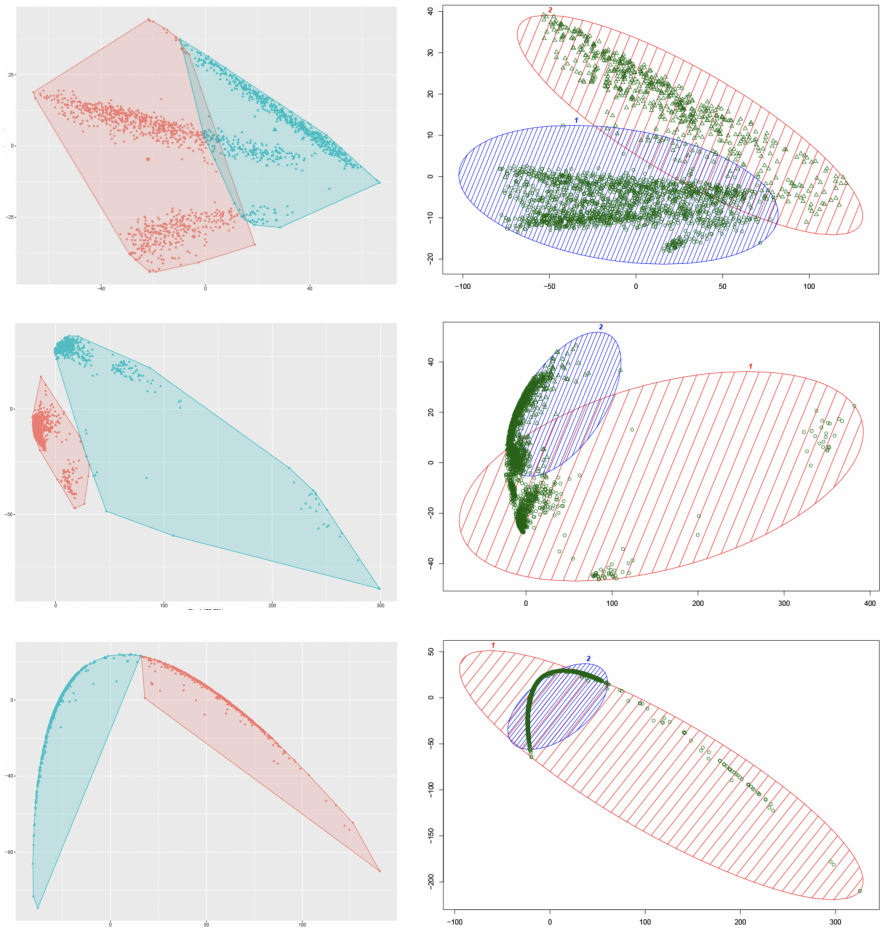


Figure 4.2: Graphical representation of the clusters obtained with K-means (LSA (a), Text2Vec (c), Doc2Vec (e)) and Fuzzy c-means (LSA (b), Text2Vec (d), Doc2Vec (f)) algorithms for each technique.

Both algorithms needs the definition of K, the number of clusters in which the

information is divided. Because we classify the patient in two groups (genetically affected, genetically not affected) we set $K = 2$. Moreover, as we have already said, in the Fuzzy c-means technique an element can belong to several clusters without restrictions and with different percentages of belonging. To overcome this problem, we have cleaned up all the spurious values, i.e., the observations that were less than $1/K$, which were therefore removed and the results obtained are shown in Fig. 4.2.

To perform the comparison of the processes, as we have already said previously, the techniques were performed on the dataset divided by visits. Considering that the visits follow a precise chronological order (starting from the SC-screening visit up to the LOG-diagnostic visit) an incremental subdivision of the dataset has been chosen, therefore the subset of the dataset relating to a specific visit will contain the data of it plus those of previous visits. Furthermore, we have chosen to implement a function that subdivides the dataset into a partition in which each subset corresponds to a visit and contains only the information relating to it; this is to allow future analyzes that focus solely on the data provided by a visit or by groups of visits.

The results of the clustering algorithms are analyzed, by computing Precision, Recall and F-measure for each cluster. Precision and Recall are measures used to indicate accuracy and completeness of results, respectively, while F-measure represents a trade-off between these measures. All the measures are based on a comparison between an expected result and the result obtained. In particular, Precision measures (also called positive predictive value) the ratio between the correctly obtained instances (true positives) with respect to the total number of instances returned by the processing process (true positive and false positive). The Recall measures (also known as sensitivity) the ratio of instance correctly obtained (true positive) with respect to the number of expected instances (true positive and true negative) [75].

$$Precision = \frac{|R \bigcap D|}{|D|}, \;\; Recall = \frac{|R \bigcap D|}{|R|} \tag{4.2}$$

The F-measure is the harmonic mean of Precision and Recall and provides a measure of how the processing is effective.

$$F - measure = 2 * \frac{Precision * Recall}{Precsion + Recall} \tag{4.3}$$

where D is the list of patients returned by the classification and R is the list of patients correctly classified. The results of this analysis process are summarized in Figure 4.3 and 4.4, where is highlighted that the best classification accuracy is reached with the fuzzy technique in each visit.

Visit	LSA K-MEANS									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	585	201	786	270	480	750	0,74	0,68	0,71	1536	855	681
BL	594	537	1131	463	390	853	0,53	0,56	0,54	1984	1057	927
V01	596	537	1133	461	390	851	0,53	0,56	0,54	1984	1057	927
V02	607	547	1154	450	380	830	0,53	0,57	0,55	1984	1057	927
V03	615	545	1160	442	382	824	0,53	0,58	0,55	1984	1057	927
V04	636	551	1187	421	376	797	0,54	0,60	0,57	1984	1057	927
V05	631	540	1171	426	387	813	0,54	0,60	0,57	1984	1057	927
V06	639	537	1176	418	390	808	0,54	0,60	0,57	1984	1057	927
V07	640	535	1175	417	392	809	0,54	0,61	0,57	1984	1057	927
V08	634	531	1165	423	396	819	0,54	0,60	0,57	1984	1057	927
V09	635	531	1166	422	396	818	0,54	0,60	0,57	1984	1057	927
V10	633	528	1161	424	399	823	0,55	0,60	0,57	1984	1057	927
V11	635	527	1162	422	400	822	0,55	0,60	0,57	1984	1057	927
V12	630	526	1156	427	401	828	0,54	0,60	0,57	1984	1057	927
V13	595	527	1122	462	400	862	0,53	0,56	0,55	1984	1057	927
V14	638	561	1199	419	366	785	0,53	0,60	0,57	1984	1057	927
V15	648	557	1205	409	370	779	0,54	0,61	0,57	1984	1057	927
ST	647	557	1204	410	370	780	0,54	0,61	0,57	1984	1057	927
LOG	556	457	1013	501	470	971	0,55	0,53	0,54	1984	1057	927

Visit	TEXT2VEC K-MEANS									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	760	205	965	96	476	572	0,79	0,89	0,83	1537	856	681
BL	960	67	1027	97	860	957	0,93	0,91	0,92	1984	1057	927
V01	749	217	966	308	710	1018	0,78	0,71	0,74	1984	1057	927
V02	757	426	1183	300	501	801	0,64	0,72	0,68	1984	1057	927
V03	765	410	1175	292	517	809	0,65	0,72	0,69	1984	1057	927
V04	1043	922	1965	14	5	19	0,53	0,99	0,69	1984	1057	927
V05	1043	922	1965	14	5	19	0,53	0,99	0,69	1984	1057	927
V06	1043	922	1965	15	5	20	0,53	0,99	0,69	1985	1058	927
V07	764	614	1378	293	313	606	0,55	0,72	0,63	1984	1057	927
V08	763	615	1378	294	312	606	0,55	0,72	0,63	1984	1057	927
V09	765	613	1378	292	314	606	0,56	0,72	0,63	1984	1057	927
V10	766	613	1379	291	314	605	0,56	0,72	0,63	1984	1057	927
V11	770	612	1382	287	315	602	0,56	0,73	0,63	1984	1057	927
V12	770	611	1381	287	316	603	0,56	0,73	0,63	1984	1057	927
V13	772	610	1382	285	317	602	0,56	0,73	0,63	1984	1057	927
V14	772	610	1382	285	317	602	0,56	0,73	0,63	1984	1057	927
V15	772	610	1382	285	317	602	0,56	0,73	0,63	1984	1057	927
ST	773	610	1383	284	317	601	0,56	0,73	0,63	1984	1057	927
LOG	766	615	1381	291	312	603	0,55	0,72	0,63	1984	1057	927

Visit	DOC2VEC K-MEANS									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	775	220	995	81	461	542	0,78	0,91	0,84	1537	856	681
BL	768	244	1012	289	683	972	0,76	0,73	0,74	1984	1057	927
V01	747	165	912	310	762	1072	0,82	0,71	0,76	1984	1057	927
V02	762	309	1071	295	618	913	0,71	0,72	0,72	1984	1057	927
V03	756	256	1012	301	671	972	0,75	0,72	0,73	1984	1057	927
V04	767	382	1149	290	545	835	0,67	0,73	0,70	1984	1057	927
V05	762	395	1157	295	532	827	0,66	0,72	0,69	1984	1057	927
V06	762	400	1162	295	527	822	0,66	0,72	0,69	1984	1057	927
V07	757	400	1157	300	527	827	0,65	0,72	0,68	1984	1057	927
V08	758	400	1158	299	527	826	0,65	0,72	0,68	1984	1057	927
V09	753	399	1152	304	528	832	0,65	0,71	0,68	1984	1057	927
V10	752	400	1152	305	527	832	0,65	0,71	0,68	1984	1057	927
V11	752	399	1151	305	528	833	0,65	0,71	0,68	1984	1057	927
V12	752	398	1150	305	529	834	0,65	0,71	0,68	1984	1057	927
V13	752	414	1166	305	513	818	0,64	0,71	0,68	1984	1057	927
V14	752	431	1183	305	496	801	0,64	0,71	0,67	1984	1057	927
V15	760	515	1275	297	412	709	0,60	0,72	0,65	1984	1057	927
ST	760	514	1274	297	413	710	0,60	0,72	0,65	1984	1057	927
LOG	761	513	1274	296	414	710	0,60	0,72	0,65	1984	1057	927

Figure 4.3: Clustering results for the K-means algorithm and Precision, Recall and F-measure data.

Visit	LSA FUZZY CLUSTERING									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	855	483	1338	0	198	198	0,64	1,00	0,78	1536	855	681
BL	563	239	802	494	688	1182	0,70	0,53	0,61	1984	1057	927
V01	563	241	804	494	688	1182	0,70	0,53	0,61	1986	1057	929
V02	560	239	799	497	688	1185	0,70	0,53	0,60	1984	1057	927
V03	624	454	1078	433	473	906	0,58	0,59	0,58	1984	1057	927
V04	626	448	1074	431	479	910	0,58	0,59	0,59	1984	1057	927
V05	628	443	1071	429	484	913	0,59	0,59	0,59	1984	1057	927
V06	621	432	1053	436	495	931	0,59	0,59	0,59	1984	1057	927
V07	622	430	1052	435	497	932	0,59	0,59	0,59	1984	1057	927
V08	624	433	1057	433	494	927	0,59	0,59	0,59	1984	1057	927
V09	628	436	1064	429	491	920	0,59	0,59	0,59	1984	1057	927
V10	634	447	1081	423	480	903	0,59	0,60	0,59	1984	1057	927
V11	634	447	1081	423	480	903	0,59	0,60	0,59	1984	1057	927
V12	635	450	1085	422	477	899	0,59	0,60	0,59	1984	1057	927
V13	773	640	1413	284	287	571	0,55	0,73	0,63	1984	1057	927
V14	509	315	824	548	612	1160	0,62	0,48	0,54	1984	1057	927
V15	221	93	314	836	834	1670	0,70	0,21	0,32	1984	1057	927
ST	221	93	314	836	834	1670	0,70	0,21	0,32	1984	1057	927
LOG	222	93	315	835	834	1669	0,70	0,21	0,32	1984	1057	927

Visit	TEXT2VEC FUZZY CLUSTERING									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	764	205	969	92	476	568	0,79	0,89	0,84	1537	856	681
BL	776	26	802	281	901	1182	0,97	0,73	0,83	1984	1057	927
V01	756	70	826	301	857	1158	0,92	0,72	0,80	1984	1057	927
V02	774	405	1179	283	522	805	0,66	0,73	0,69	1984	1057	927
V03	772	405	1177	285	522	807	0,66	0,73	0,69	1984	1057	927
V04	776	414	1190	281	513	794	0,65	0,73	0,69	1984	1057	927
V05	777	421	1198	280	506	786	0,65	0,74	0,69	1984	1057	927
V06	778	419	1197	279	508	787	0,65	0,74	0,69	1984	1057	927
V07	777	414	1191	280	513	793	0,65	0,74	0,69	1984	1057	927
V08	777	413	1190	280	514	794	0,65	0,74	0,69	1984	1057	927
V09	776	409	1185	281	518	799	0,65	0,73	0,69	1984	1057	927
V10	749	94	843	308	833	1141	0,89	0,71	0,79	1984	1057	927
V11	744	63	807	313	864	1177	0,92	0,70	0,80	1984	1057	927
V12	740	55	795	317	872	1189	0,93	0,70	0,80	1984	1057	927
V13	739	50	789	318	877	1195	0,94	0,70	0,80	1984	1057	927
V14	736	43	779	321	884	1205	0,94	0,70	0,80	1984	1057	927
V15	736	43	779	321	884	1205	0,94	0,70	0,80	1984	1057	927
ST	735	43	778	322	884	1206	0,94	0,70	0,80	1984	1057	927
LOG	735	44	779	322	883	1205	0,94	0,70	0,80	1984	1057	927

Visit	DOC2VEC FUZZY CLUSTERING									Patients		
	Cluster PD			Cluster GP			Results					
	PD	GP	Total	PD	GP	Total	Precision	Recall	F-measure	Total	PD	GP
SC	518	455	973	337	224	561	0,53	0,61	0,57	1534	855	679
BL	754	456	1210	303	471	774	0,62	0,71	0,67	1984	1057	927
V01	735	426	1161	322	501	823	0,63	0,70	0,66	1984	1057	927
V02	838	376	1214	219	551	770	0,69	0,79	0,74	1984	1057	927
V03	811	364	1175	246	563	809	0,69	0,77	0,73	1984	1057	927
V04	931	456	1387	126	471	597	0,67	0,88	0,76	1984	1057	927
V05	918	435	1353	139	492	631	0,68	0,87	0,76	1984	1057	927
V06	934	446	1380	123	481	604	0,68	0,88	0,77	1984	1057	927
V07	933	438	1371	124	489	613	0,68	0,88	0,77	1984	1057	927
V08	925	428	1353	132	499	631	0,68	0,88	0,77	1984	1057	927
V09	924	419	1343	133	508	641	0,69	0,87	0,77	1984	1057	927
V10	923	421	1344	134	506	640	0,69	0,87	0,77	1984	1057	927
V11	921	422	1343	136	505	641	0,69	0,87	0,77	1984	1057	927
V12	924	424	1348	133	504	637	0,69	0,87	0,77	1985	1057	928
V13	925	423	1348	132	504	636	0,69	0,88	0,77	1984	1057	927
V14	926	422	1348	131	505	636	0,69	0,88	0,77	1984	1057	927
V15	930	421	1351	127	506	633	0,69	0,88	0,77	1984	1057	927
ST	930	421	1351	127	506	633	0,69	0,88	0,77	1984	1057	927
LOG	934	408	1342	223	519	742	0,70	0,81	0,75	2084	1157	927

Figure 4.4: Clustering results for Fuzzy c-means algorithm and Precision, Recall and F-measure data.

Specifically, as shown in Fig. 4.5, from the comparison of the results, it is shown that the techniques to which the k-means clustering is applied produce lower results than the Fuzzy clustering, except for the Doc2Vec in the SC, V02, V03 visits and the Text2Vec in the SC, BL, V01 visits where they produce better results than the other algorithms. Furthermore, despite the excellent performance of the Text2Vec k-means in the BL visit (F-Score 0,92), the technique proves to be somewhat unstable, in fact, we can see its total failure in the V04-V06 visits, where the k-means fails creating a cluster of a few dozen items. While the Text2Vec technique with Fuzzy clustering is the process that produces the best results in the series of visits from SC to V02 and from V10 to LOG. During the screening visit, both Doc2Vec (k-means) and Text2Vec (k-means and Fuzzy) produce very precise and similar classifications of patients. From what has been observed we can conclude that the processes with LSA produce the worst results, and generally, the results produced by the execution of the processes via K-means are lower than those with Fuzzy clustering, this is due to the intrinsic structure of the K-means algorithm that does not allow an element that can be positioned in both clusters or that it can be moved later from one cluster to another, forcing an incorrect classification of the information processed. This is because in our data there is the PRODROMAL class, which are subjects not affected by Parkinson Disease but who present symptoms characteristic of this disease, therefore a stringent clustering could lead to a higher error rate, as can be seen from the results. In conclusion, the techniques that currently produce the best results are Text2Vec-Fuzzy, in visits from SC to V02 and from V10 to LOG, and Doc2VecM-Fuzzy in visits from V03 to V09.

Finally, we calculated the frequency of the different terms present in the documents and we have discarded the 95 quantile. we have extracted the most relevant words from the documents related to the patient in the LOG visit, than we classified them in five categories:

- *Parental*: Patern and Sibling;
- *Symptomatic*: Pain, Sleep, Thyroid, Muscoloskeletal and Urinary;
- *Related disorders*: Hypothyroidism, Hypercholesterolemia, Diabet and Reflux;
- *Therapeutic*: Amantadin, Rytary, Arilict, Risagilin, Mirapex and Levodopa;
- *General terms*: Procedure and Full.

The words that are of most interest are Hypothyroidism, Hypercholesterolemia and Diabet. Regarding hypothyroidism, although no evidence of a higher frequency of hypothyroidism among patients with Parkinson's disease has been reported in the

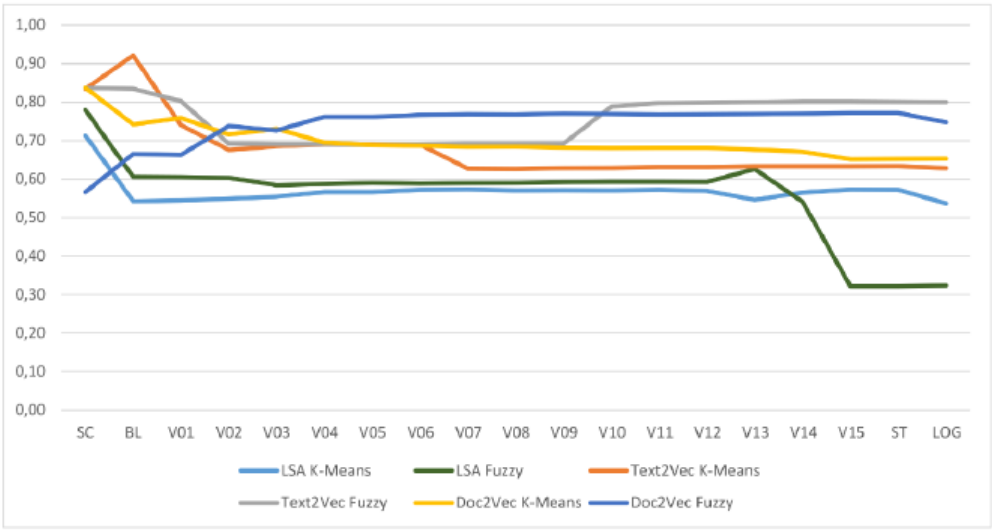


Figure 4.5: Comparison among F-measures for k-means and Fuzzy c-means algorithms.

literature, there may be a concomitance between these two diseases. In fact, studies conducted on patients with PD who take levodopa / carbidopa have indicated that the reduction of TSH levels is directly related to the drug, and occurs only during the first two hours after intaking. It is not related to any significant thyroid dysfunction. This effect tends to be more discernible in males and is probably related to a primary or secondary propensity to hypothalamic levels specific for patients with PD [162]. Another study, however, has shown that the thyroid gland and its enzyme thyroperoxidase participate in the nitrosylation of serum proteins and can influence parkinsonian nitrosative stress and nitrosylation of serum alpha-synuclein, a potentially pathogenic facto [76]. Concerning hypercholesterolemia, many studies have not found a close correlation between the two pathologies, only a large prospective study [108] suggests that high total cholesterol at baseline is associated with an increased risk of Parkinson's disease. As for diabetes, a large study [240] showed that diabetes was associated with a higher future risk of PD, because the insulin receptors are expressed in the substantia nigra. The dopamine agonist bromocriptine improves glycemic control and was approved for adjunctive treatment of diabetes. Conversely, the insulin sensitizer rosiglitazone protects dopaminergic neurons in animal models of PD. It is also important to point out that both diabetes and PD are age-related chronic diseases and some pathogenic processes may underlie both

conditions.

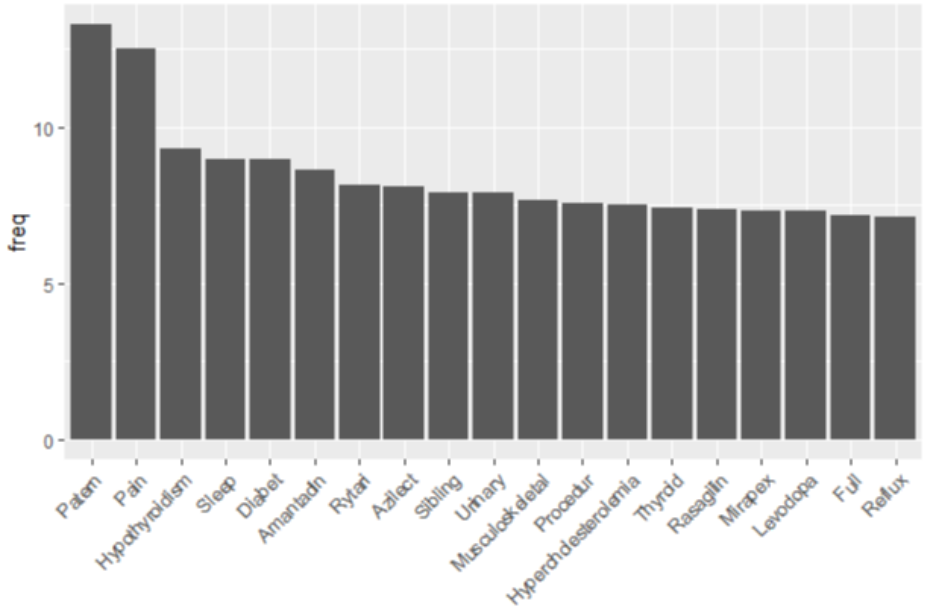


Figure 4.6: Barplot representing the words with the highest frequencies in the documents related to the LOG visits.

The visualization of the information is very important in each field for a personalized visualization and a better understanding of the information but especially in the medical field. For example, it can act as a decision and diagnosis support for clinicians. There are many tools for viewing information and data, here we use the D3.js framework[1] that is a JavaScript library to create dynamic and interactive visualizations starting from organized data, visible through a common browser. for the visual rendering of the data, to be able to create both small tables, diagrams and statistics and complex graphic representations (including animations and other possibilities of interaction). Libraries are always linked to software that uses the functions of a programming library when a specific function of the collection is requested, which is why they only work within a program and cannot be performed independently.

In our case, we offer a visualization system, based on Radar chart, useful for the clinician to place each individual patient in one of the two clusters, in order to be able to make a quicker and faster diagnosis as the visits made in chronological order follow one another. For this visualization, we used the data obtained from both

[1] https://d3js.org

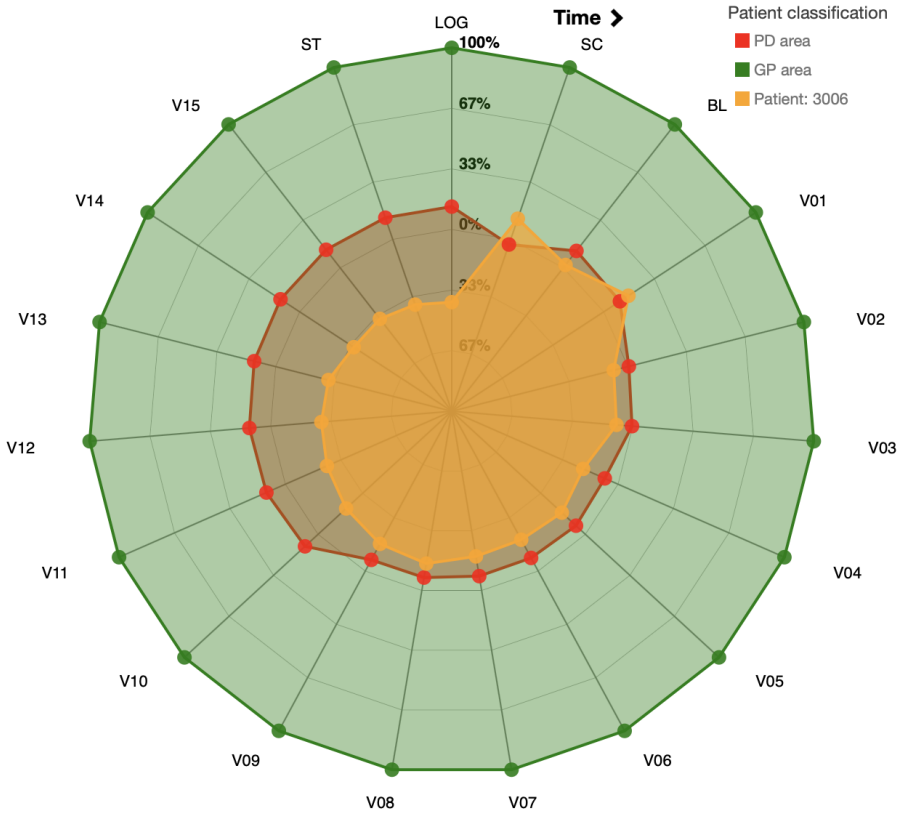


Figure 4.7: A data visualization based or Radar chart for fast diagnosis.

Text2Vec and Doc2Vec, based on the Fuzzy clustering efficiency obtained from the processing carried out for each visit. In fact, we used the Text2Vec for the SC, BL, V01, V02, V10-V15, ST and LOG visits and Doc2Vec for all the other visits, thus obtaining the chart shown in Fig. 4.7, for patient 3006. The chart is composed of three areas the green one indicates the non-sick patient, while the red one the sick patient; the orange line indicates the probability of patient 3006 to be included in the PD and GP areas. In the V02-V03 visits, the values indicate that the patient in question can be affected by Parkinson's disease. Unfortunately, this is confirmed more and more in subsequent visits.

4.2 AI for supporting the detection of Coloboma Disease

In this section I present the results of my research iris recognition when it is affected by Coloboma, In particular, I describe how the Daugman and Canny edge detection algorithms fails to detect irises with this pathology. I propose an AI methods for detecting this disease and extend the Daugman algorithm to recognize also people affected by Coloboma.

4.2.1 What is Coloboma?

The Coloboma disease is a congenital malformation characterized by the absence of a portion of the iris, which is not perfectly welded during embryonic development. In ophthalmological use, the word refers to any notch, gap, hole or fissure in any of the congenital or acquired ocular structures; for this reason, this anomaly can be described as a “hole-shaped crack lock" [172]. This congenital anomaly can also be caused by genetic factors and implies the absence of tissue in one or more ocular structures, such as the cornea, iris, ciliary body, crystalline, retina, choroid and optic disk.

The Coloboma disease can be classified as typical or atypical, depending on its position in the iris or fundus [172]. This defect is found from 0.5 to 0.7 individuals in about 10000 births. Such a disease can affect one or both eyes (unilateral or bilateral Coloboma). The spillovers on vision depend on the location and extent of the ocular malformation.

4.2.2 Dataset

We collected images of 238 people’s eyes: 38 suffering from Coloboma, collected by searching on Google images with the keywords "iris malformation" and "Coloboma", 10 belonging to African-Americans people (due to low contrast between iris and pupil) [146]. searched with the same keywords plus "African-Americans", and 190 belonging to Caucasians, randomly taken from the dataset Casia.

In order to equilibrate the number of images in each individual category and to improve the image quality we conducted on the dataset a *data augmentation* and a *pre-processing* phase, respectively.

Data augmentation is a technique to create new data from existing training data using various transformations [130], including translation, rotation, flipping, cropping, adding noises [180]. This phase is performed to reduce the data-imbalance issue by increasing the size of the dataset with respect to the 38 images of eyes suffering from Coloboma and the 10 ones belonging to African-Americans. Specifically, the following geometric-based transformations have been adopted, constrained to

preserve the continuity of the lines and the reciprocal relationships of position and proportion of the image elements:

- *flip*: for each image in the data collected, we performed first a horizontal flipping of pixels around the horizontal axis, and then we reversed the pixels horizontally;
- *rotation*: for each image in the data collected, we created three new images by applying a rotation transformation using rotation angles $-20°$, $20°$, $180°$and $200°$ with respect to the horizontal axis.

The resulting dataset is composed by 478 iris images (see Table 4.1).

Label	elements	Description
0	250	*Non-pathological iris*
1	228	*Coloboma iris*

Table 4.1: The dataset composition

Then, we used the pre-processing to reduce the image pattern complexity and consequently increase the performance achieved in the experiments, we performed the following steps:

a) each pixel in the image represented by a RGB coefficient in the integer range $[0, 255]$ was scaled in the floating-point range $[0, 1]$ by applying the $1/255$ factor;

b) we converted image into grayscale;

c) we applied the *morphological closing operator* to remove small flaws by preserving the shape and size of the objects in the image [132];

d) to reduce noise in image, we used the *median filter* where the value of each pixel in the output image is the median of surrounding neighbourhood [239];

e) to increase the image contrast, we used *contrast-limited adaptive histogram equalization* [209].

f) we used *Otsu's threshold*to mitigate the reflections which often are in pupils. Otsu's method is adopted to automatically determine the contours of the image by reducing it from a grayscale to a binary image with respect a computed threshold [223]. Indeed, Otsu's method assumes that the image contains two classes of pixels based on the bimodal histogram, and calculates the optimum threshold by separating the two classes by minimizing their intra-class variance, and maximizing the extra-class one.

4.2.3 On the Limitation of Pathological Iris Recognition: Neural Network Perspectives

First of all, we used Daugman's algorithm, that consists of two steps: *(1)* pupil and iris detection, and *(2)* iris recognition. During the detection phase, the main problem is the recognition of the margins of a pupil inside an image. This step is crucial because it ensures that the same coordinates are assigned to certain parts of the iris in different images. The basic idea is that the pupil appears darker than the iris, with a drastic brightness change between iris and pupil. Assuming that the pupil has a circular shape, a series of concentric circumferences is traced. By adding up the brightness that each circumference has in turn for each point, the maximum value is reached when matching the centre and radius of the pupil. In this way it is possible to find the iris's edge, delimited by the sclera.

The algorithm is particularly efficient when the iris is visible and no part of it is covered by the eyelid, which could be mistakenly identified as the edge of the iris. To overcome this issue, a procedure similar to that used for the iris is adopted to localize the upper and lower eyelid line. The algorithm maps each iris point into a point in polar coordinates (r, θ), and adapts to the dilation and variability in dimension of the pupil, by producing an invariant representation with respect to size and translation in the polar coordinate system. During the recognition phase, the rotation is considered by translating the iris template in the direction θ until the two irises are aligned to be compared. Because the radial coordinate varies from the internal contour of the iris to the external one in a unitary interval, the Daugman's algorithm performs an intrinsic correction of the elastic deformation of the iris due to dilation/contraction of the pupil. The pupil and iris edges are extracted by applying the *Gabor filter* to the image [55] in the polar coordinate system. The filter produces as output a code of 256 bytes (2048 bits).

The comparison between two iris codes is based on the *Hamming distance* between the two images 256-byte codes. If the two codes are generated by the same iris, the Hamming distance will be close to 0, due to the high correlation. Two images are said to be independent if their fractional Hamming distance is above the 0.33 threshold [56].

To manage possible rotations, the codes are one-bit shifted to the left and the right and therefore more Hamming distances are calculated accordingly. For recognition, the lowest Hamming distance is selected as it corresponds to the best match between the two templates.

Similarly we used Canny algorithm that is an edge detector with a low error rate, which means it can capture most of the edges with high precision by recognizing as many real edges as possible within the image. It is mainly used for iris detection and is composed by four main stages:

1. *Noise reduction.* A primary problem in the contour recognition algorithms is

the presence of noise in the unprocessed images. So as a first step a spatial filter is applied to the image. The aim is to remove the high frequencies where the noise interference is more problematic. A *Gaussian filter* is chosen because its standard deviation value can be modified to favor the recognition of small and sharp edges or larger and gradual edges. The result of this first phase is a slightly blurred image, in which no pixel is significantly affected by noise.

2. *Search for the gradient of image brightness.* Two gradients are calculated on the x and y axis by using the *Sobel operator* that returns the value of the first derivative of both the horizontal and vertical direction. This operator guarantees a fair approximation while keeping the calculation request low. Then, for each pixel the direction angle θ between gradients is computed and rounded to the values of 0, 45, 90, and 135, representing the vertical, horizontal, and the two diagonal directions, respectively.

3. *Elimination of non-maxima.* In this phase, the values of the pixels not considered part of the outline are reset, i.e., the pixels whose intensity value is not greater than that of the adjacent pixels located along the direction given by the value θ at that point. The result is a binary image with a thin line at the edges of the objects in the image. The algorithm compares the intensity of the current pixel border with the intensity of the pixel border in the positive and negative gradient directions. If the intensity of the edge of the current pixel is greater than the other pixels of the mask in the same direction, the value will be preserved, otherwise, the value will be suppressed.

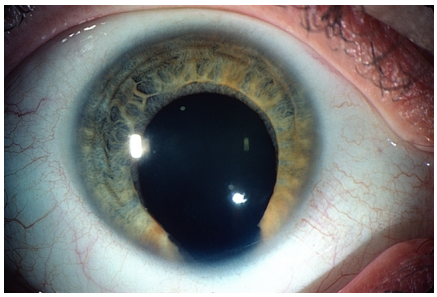


Figure 4.8: An eye suffering from Coloboma disease.

4. *Identification of the contours by thresholding with hysteresis.* A weakness concerning the contour recognition algorithms is given by the threshold value selected to obtain a good result. This value is often determined empirically, by predicting what is the minimum value that the gradient must-have for the considered pixel to be part of the outline. To mitigate this weakness, the threshold-

ing with hysteresis method is used. This method is based on the high threshold value and the low threshold value. Each point whose gradient is higher than the high threshold is automatically defined part of the contour (value 1). Furthermore, each point contiguous to a point of the contour that has a gradient value higher than the low threshold joins the contour. Finally, a value of 0 is assigned to all the remaining points and those below the low threshold. The result is the binary image of the outlines.

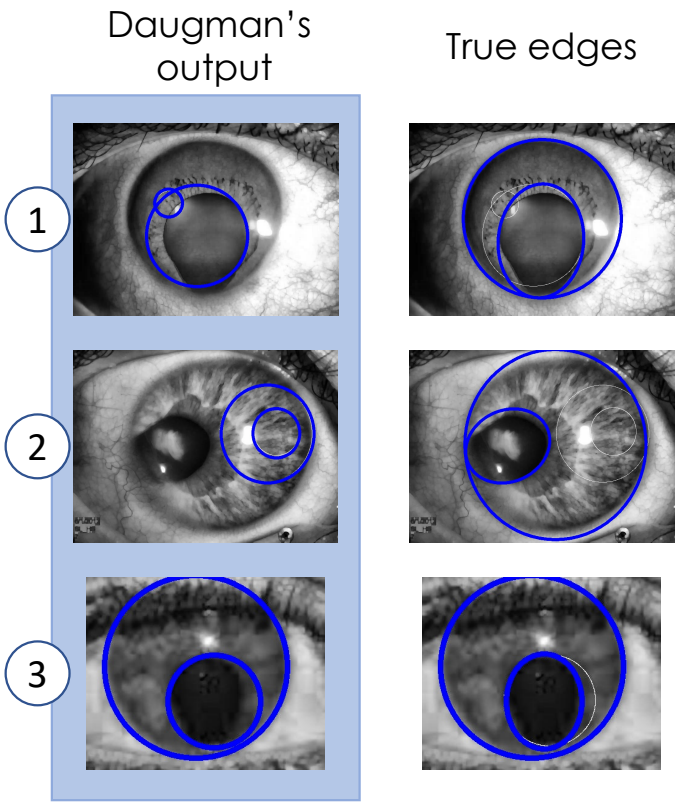


Figure 4.9: Visualization of errors made by Daugman's algorithm (left) versus true pupils and iris edges (right).

Then, we show the Dougman and the Canny edge detection algorithms inability in dealing with iris suffering from Coloboma and their performance degradation. Observe that irises suffering from such pathological state are not perfectly concentric to the pupils: the radial distance from the right side of the limit between sclera

and iris, and the left side is not always the same but may vary.

We applied these two algorithms to the 38 irises suffering from Coloboma in the data collected as detailed in Section **??**. In the left-hand side of Fig. 4.9 we show the wrong edges (pupil and iris) returned by Daugman's algorithm concerning irises suffering from Coloboma in our dataset; on the right side the true edges (hand-traced). Specifically, for 5 images (13.16% of the number of Coloboma eye image in the dataset) the algorithm returned a non-image acquisition error, i.e., it failed in iris detection because it was unable to perform the segmentation task. For 6 images (15.79%) the algorithm correctly (although often not perfectly) detected the pupil and the iris (see part 3 in Fig. 4.9). For the remaining 27 images (71.05%), the algorithm completed the iris detection, but the segmentation was wrong (see parts 1-2 in Fig. 4.9).

When executing the Canny edge detection algorithm on our Coloboma images we obtained the non-acquisition of the pupil edge for 20 images (52.63%) (see part 3 in Fig. 4.10), while for 18 images (47.37%) we observed a good detection of both the edge of the iris and the pupil (see parts 1-2 in Fig. 4.10).

In order to measure errors obtained by the Daugman algorithm and Canny algorithm. We calculated the *edge distance* between the circular edges of the iris and pupil, as detailed in the following.

We assumed true edges are ellipse-like and circular-like shape for pupils and irises respectively. Specifically, first we traced the circular edges of the iris and pupil given by Daugman and Canny algorithms, named I', P', and the true edges of the iris and pupil, named I, P, as show in Fig. 4.9 and 4.10. Then, we defined the edge distances $D(I, I')$ ($D(P, P')$, resp.) between I and I' (P and P', resp) as follows.

$$D(I, I') = \frac{area(I) + area(I') - (2 * (area(I \cap I')))}{area(I) + area(I')}$$

$$D(P, P') = \frac{area(P) + area(P') - (2 * (area(P \cap P')))}{area(P) + area(P')}$$

where the $area$ function computes the number of pixels of the input edge, ellipse-like for P and P' and circular-like for I and I'. The distance values range in $[0, 1]$, where 0 means that the edges are perfectly overlapped, and 1 means a not intersecting edge.

For the Daugman's algorithm, we calculated this distance for the 71.05% of the images, those for which segmentation was obtained (see Fig. 4.11(a)); for the Canny edge detection algorithm we calculated the distance on the 47.37% of the images, for which we obtained good contours of both the iris and the pupil, as show in Fig. 4.11(b).

In Fig. 4.11, we plot the error in terms of edge distance made by the Daugman and Canny algorithms in the case the segmentation process has been completed (in blue is the edge distance on iris detection and in red is the one on pupil detection).

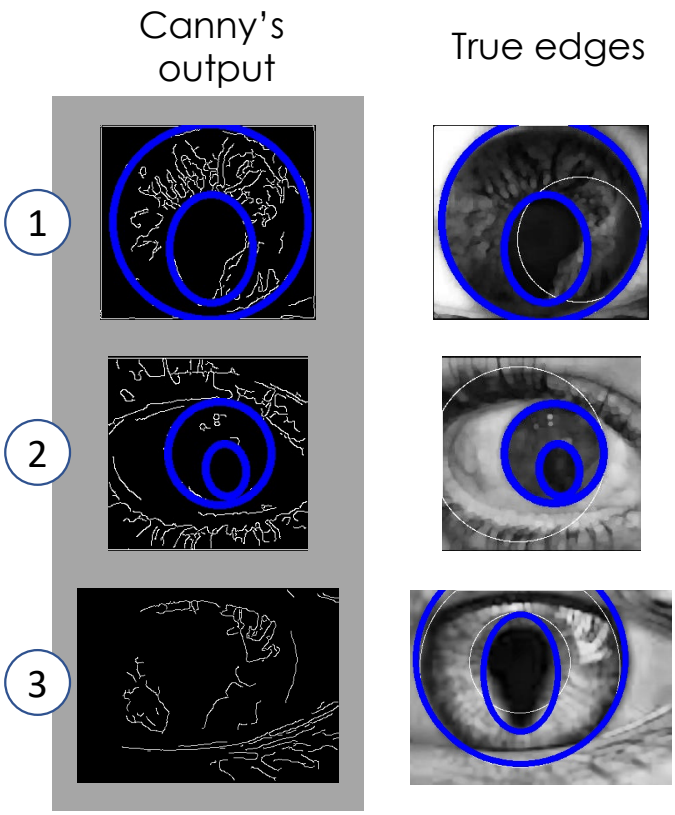
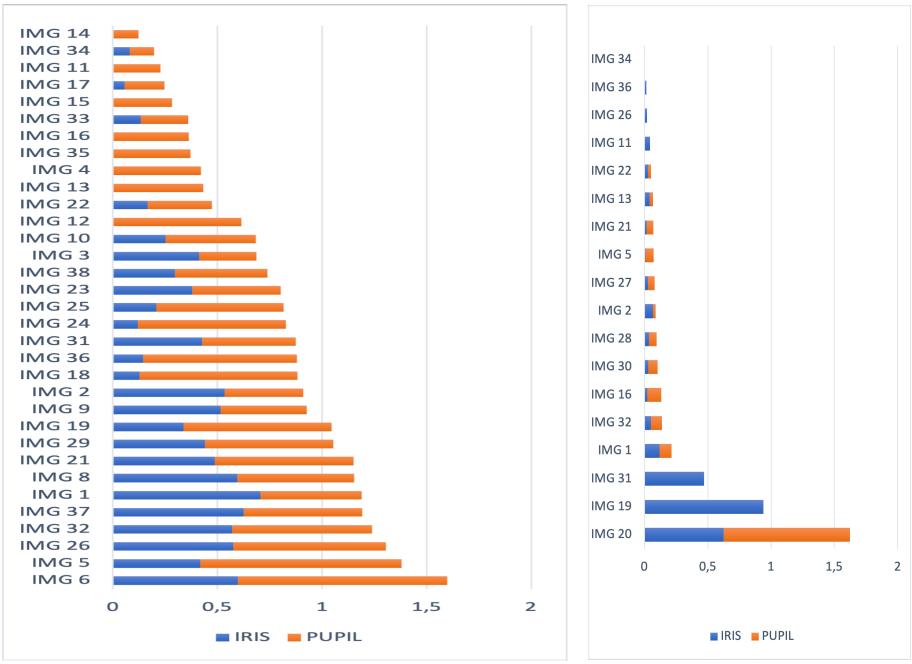


Figure 4.10: Visualization of errors made by Canny edge detection algorithm (left) versus true pupils and iris edges (right).

Based on these results, we could say that Daugman's algorithm is more robust given that it is able to segment more images but it is less precise due a higher error rate, while the Canny edge detection algorithm is less robust but it is more precise, since it segments more correctly the edges of the iris and pupil at the end of the segmentation process. However, Daugman and Canny edge detection algorithms were able to correctly detect iris and pupils only for 6/38 and 18/38 images, respectively.

In the [79] we used a Convolutional Neural Networks (CNN) for the classification of irises suffering from Coloboma. We remark that such a model was chosen for its great performance and wide use in image and disease classification (e.g., [141, 245]). Several recent works in literature have faced the problem of classifying diseases, like us for the Coloboma pathological state. For instance, in [245] authors focused on



(a) (b)

Figure 4.11: Error values between (a) the Daugman output ((b) the Canny output) and the true edges of iris and pupil.

Chronic Obstructive Pulmonary Disease diagnosis. They designed and developed a neural network that is capable to get the disease patterns, and complies with the intrinsic structure of neuromorphic platforms, addressing data safety concern. While, authors in [141] designed a simple and effective CNN with atrous convolutional layers for Pulmonary Tubercolosis Diagnosis in radiographs capable to achieve up to 96% accuracy. Observe that although we propose a CNN to detect the presence of a disease, our main goal is to move towards a solution that tackles biometric-driven discrimination by embedding our CNN in a more sophisticated biometric system.

Design: The CNN developed is structured in: 4 *Convolutional blocks*, 1 *Flattering layer*, and 1 *Dense layer* (as shown in Fig. 4.12).

Convolutional block. It consists of one *Convolutional layer* followed by one *Max-Pooling layer*:

1. *Convolutional layer*: its objective is to extract the *convolved features*, such as curves, angles, circumferences or squares depicted in an image with high pre-

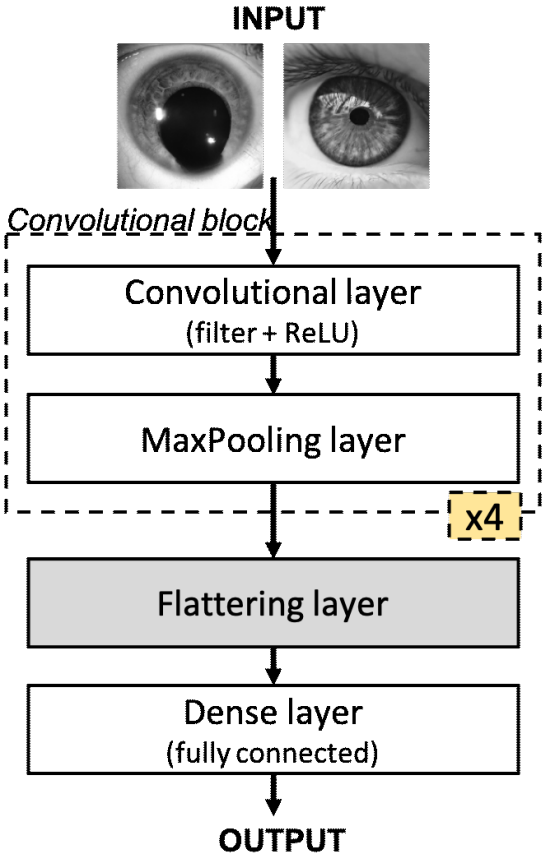


Figure 4.12: CNN to classify iris suffering from Coloboma. The CNN is structured in: 4 *Convolutional blocks*, 1 *Flattering layer*, and 1 *Dense layer*.

cision, in order to divide the image into several overlapping fragments, and then saving them as an array. This layer extracts only the significant parts of the input image by applying the *convolution kernel* (filter), which are then returned as input to the next step in the form of a feature map. The connections between nodes belonging to the convolutional layers are assigned by weighting filters having a kernel size of (K_x, K_y). In our case, due the low complexity of the processed images, we set $K_x = K_y = 3$. The convolution occurs inside the image border, and so the feature map is reduced by the $(K_x - 1, K_y - 1)$ pixels from width and height [126].

2. *MaxPooling layer*: this layer is responsible for reducing the spatial size of the convolved feature. The objective is to decrease the computational power required to process the data through dimensionality reduction. Furthermore, it is useful to extract dominant features which are rotational and positional invariant, meanwhile preventing overfitting. This layer returns the maximum value from the portion of the image covered by the kernel.

We tested from 1 to 6 Convolutional blocks and we found that 4 blocks was the best configuration for our problem.

Flattering layer. It transforms the output of the Max Pooling layer into the input to the Dense layer for the classification.

Dense layer. It represents the final level of our CNN. The neurons of such a layer are connected to all the ones of the previous level (fully connected), and have the purpose of performing high-level operations leading to the actual CNN output. Due the nature of our problem (binary classification) and specific dataset, we used one only Dense layer.

Validation and testing: we split the dataset into 80% for training and 20% for testing. We then used 20% of the training set as validation set. We searched for the best value of epochs parameter in the range $[1, 120]$ and between *ReLU* and *elu* activation functions. As a result, we obtained that 80 was the best value for training our CNN, with *Re*ctified *L*inear *U*nit activation function (*ReLU*). Furthermore, to prevent overfitting, after the last Max Pooling layer we used a drop-out layer: we tested in the range $[0.05, 0.25]$, finding that the best one is 0.25. We remark that the CNN has been tested exclusively on the original images of irises suffering from Coloboma. To evaluate our model we computed the precision, recall and F-measure [106].

As shown in the bottom part of Fig. 4.13, we obtained an accuracy of 100% on the training set, and 94% on the validation set. Lastly, in the testing phase we obtained 95.45% of accuracy on the test set, with a loss value of 0.031 (top part of Fig. 4.13). We observed that just for one image out of 38 the prediction fails, given the early stage of the Coloboma disease. The precision, recall and F-measure values were 0.875, 1.00 and 0.93, respectively.

Instead in [80] Residual Neural Networks (ResNet) for the classification of iris suffering from Coloboma. In the literature, there are few cases [26, 154, 166] in which a ResNet has been used for iris recognition.

ResNet is a particular type of CNN, which uses an innovative type of block, the Residual block, and exploits the concept of Residual learning [102]. While in a classic deep convolutional neural network several layers are stacked and trained for the task at hand, in residual learning the network tries to learn some characteristics by using a shortcut connections level. Furthermore, ResNet are easier to optimize and can acquire greater accuracy. The Residual block can provide an input to the next sequences of convolution-ReLU-convolution operations. This connection, named

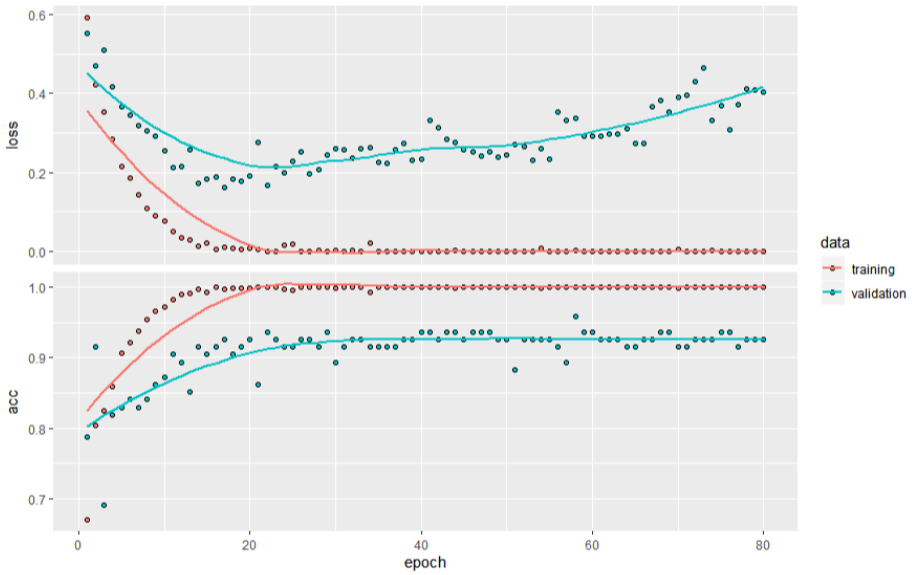


Figure 4.13: Loss and Accuracy curves over the training epochs.

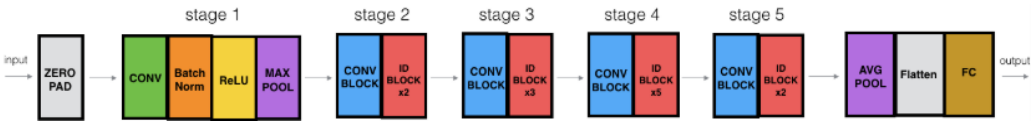


Figure 4.14: ResNet to classify iris suffering from Coloboma.

skip connection, skips one or more layers by performing a mapping identity in order to mitigate the problem of the vanishing/exploding gradient. The addition of these connections in a network neither implies the use of extra parameters, nor increases their computational complexity. The whole network can be trained by the traditional method of gradient descent with back-propagation of the error. Through the concatenation of several blocks of this type, ResNet learns to predict a certain output not by learning a direct transformation from the input data to the output, but by learning a certain term to be added to the input data to get to the output by minimizing the error, which is called residual error. Thanks to the addition operation, the back-propagation of the gradient allows to contrast the problem of gradient degradation (a problem that occurs markedly in networks with many levels slowing down the training). Another component widely used in ResNet is the batch normal-

ization level, used after each convolution and activation. Batch normalization is an operation that allows us to normalize the data present in mini-batches, and thanks to this reduces the limitations on the value of the learning rate that typically exists in the training of deep neural networks. It also makes the initialization phase of the weights less complex. As a result, the network training process time is significantly reduced.

ResNet-50 composed by 50 layer is largely adopted in the literature for image recognition [26, 218]. The adopted ResNet-50 model is shown in Fig. 4.14, and is composed as follows:

- *Zero-padding:* it pads zeros in the input to get symmetric width and height of the image.

- *Stage 1:* it performs first a 2D convolution operation (with 64 filters, a shape of $(7,7)$, and a stride of $(2,2)$ and then a Batch Normalization operation. Finally, a *ReLu* Activation and a Max Pooling operation (with a window of size $(3,3)$ and stride of $(2,2)$) are applied.

- *Stage 2:* it uses a Convolutional block with three sets of filters of size $[64,64,256]$, and then two Identity blocks with three sets of filters of the same size.

  The Convolutional block consists of layers of convolution and aims at extracting the *convolved features*, such as curves, angles, circumferences or squares depicted in an image with high precision, in order to divide the image into several overlapping fragments, and then saving them as an array. Each fragment maintains the same position it had in the original image. This layer extracts only the significant parts of the input image by applying the *convolution kernel* (filter), which are then returned as input to the next step in the form of a feature map. The result will be a set of arrays smaller than the original image, but in which the loss of information is minimal.

  In the Identity block the skip connection skips over 3 hidden layers. In the case the number of inputs is greater than the output one, padding zeros are added, otherwise a cropping operation is carried out.

- *Stage 3:* it uses a Convolutional block with three sets of filters of size $[128,128,512]$, and then three Identity blocks with three sets of filters of the same size.

- *Stage 4:* it uses a Convolutional block with three sets of filters of size $[256,256,1024]$, and then five Identity blocks with three sets of filters of the same size.

- *Stage 5:* it uses a Convolutional block with three sets of filters of size $[512,512,2048]$, and then two Identity blocks with three sets of filters of the same size.

- *Avg Pooling layer:* this layer is responsible for reducing the spatial size of the convolved feature. The objective is to decrease the computational power required to process the data through dimensionality reduction. Furthermore, it is useful to extract dominant features which are rotational and positional invariant, meanwhile preventing overfitting. This layer returns the average value from the portion of the image covered by the kernel. It has been used with window of shape $(2, 2)$.

- *Flatten layer:* it transforms the output of the Avg Pooling layer into the input to the Dense layer for the classification.

- *Dense layer:* it represents the final level of the neural network. The neurons of such a layer are connected to all the ones of the previous level (fully connected), and have the purpose of performing high-level operations leading to the actual output of the ResNet. It has been used with a *softmax* activation parameter.

We split the dataset into 90% for training and 10% for testing. We remark that the test set considered only images not produced by the data augmentation phase.

For the evaluation of the proposed model we adopted the accuracy metrics [207]. Informally, the accuracy is the number of predictions that our model predicted correctly.

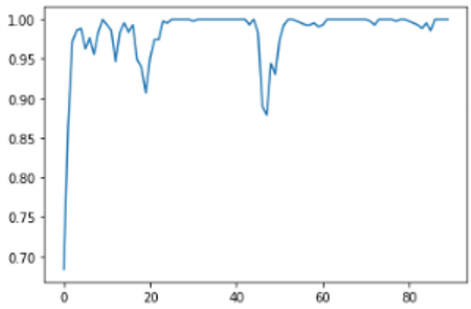
Let tp, fn, fp, and tn be the number of true positives, false negatives, false positives and true negatives, respectively.

$$accuracy = \frac{(tp + tn)}{(tp + tn + fp + fn)}$$

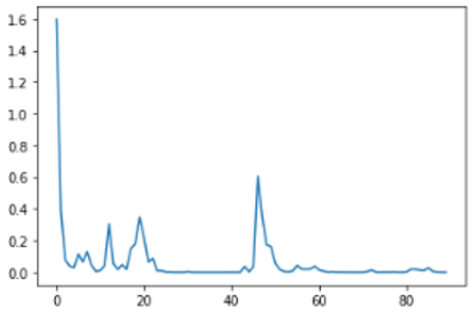
We searched for the best value of accuracy when the epochs parameter is in the range $[70, 100]$. As a result, 90 was the best value for training our ResNet-50, as shown in Fig. 4.15(a). We used $adam$ as optimizer, set $batch\_size$ equal to 32, and calculated the loss by using the $categorical\_crossentropy$.

ResNet-50 let us reach get an accuracy of 99.79% in the testing phase with a loss value of 0.000299 (see Fig. 4.15).

The results provided by the ResNet-50 have produced an improvement of 4.53% and are very promising. This highlights the possibility of using ResNet-50 for an accurate classification of the Coloboma pathological state. This means that we are able to recognize with high accuracy the iris pathological state which causes the detection problem to the most accurate algorithms. Any biometric recognition system should include a preliminary phase in which the presence of any pathology should be detected and then an appropriate customized biometric technique should selected.



(a)



(b)

Figure 4.15: Accuracy (a) and Loss (b) curves over the training epochs.

4.2.4 Extension of the Daugman algorithm

From the results of the previous section, a person affected by Coloboma may be excluded by the access to services secured by iris recognition. To avoid this inconvenient I decided to extend the Daugman's algorithm in such a way to be able to recognize also people with this pathology.

The idea behind the process is to simulate the effects of the Coloboma, obtaining data that are as consistent as possible with the original ones, through a pixel resize. The original pupil is manipulated to get an elliptical shape, expanding mainly in three directions affected by Coloboma: right, left, and down. At the same time, there is a compression of the internal region of the iris, which is not overwritten by the pupil, so that its characterizing information is not lost. Indeed, its external contours remain unchanged, without expanding them on the surface of the sclera. For each of the images used, the pupil resizing process is then carried out in each of the three directions, up to the limit of the iris. At this point, to obtain more data available, and above all to have the possibility of testing the segmentation algorithm in cases where the pupil covers almost the entire iris, a total resize of the pupil is performed

until progressively the limit of the iris is reached. Iteratively, the process similarly proceeds with each of these images in the three directions.

In the following first I describe the processing steps needed for creating a synthetic dataset, then the extension of the Daugman's algorithm I propose.

Syntetic dataset creation

The following two phases are conducted to create a synthetic dataset of eyes simulating the presence of Coloboma disease.

- *Creating the Masks and Boundaries*

  The first phase includes the formation of two manually traced truth masks, one for the iris and one for the pupil. Figure 4.16 shows on the left-side the iris mask, while the pupil mask is shown on the right. The latter, in particular, will be used as a starting point to automatically create all the other truth masks of the modified images, whose dimensions will be calculated starting from the relative scaling parameters. The iris mask, on the other hand, will always remain the same for each image, to emphasize its stability.

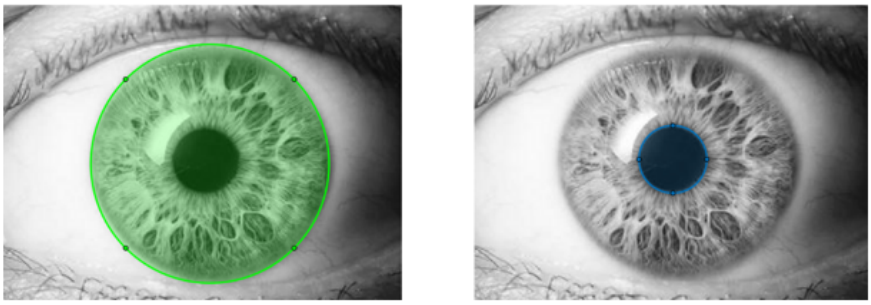


Figure 4.16: Process of creating masks.

Starting from the masks, two matrices are created, one for the rows and one for the columns. They store the start and end coordinates, expressed in pixels, of the iris and pupil, respectively. This information will be of particular importance when the pupil is resized, to ensure that the iris compressed remains confined to its original surface, without ever overstepping its limits. The limits of the iris and pupil are sufficient when resizing in a single direction of the pupil; but with the successive need to grow the pupil in its entirety, additional information is needed to identify the parts of the iris that are not aligned with the pupil in one of the four direction and which in any case need a diagonal

compression. Therefore, there is the need of a double translation of the pixels in two directions at the same time, to avoid a loss of information that would be caused by their overwriting. The intersections are calculated simultaneously with the boundaries, as shown in Figure 4.17. It is also necessary to take into account the four "corners" that do not have pixels aligned with the pupil, neither horizontally nor vertically, and that in the resize phase require compression independent of pupil growth.

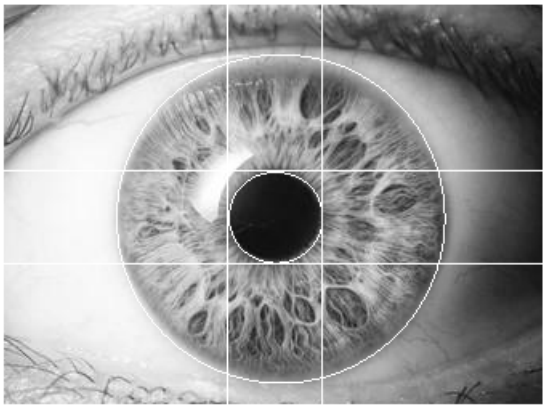


Figure 4.17: Boundaries ed Intersections.

- *Resize of the Pupil* The mask obtained for the pupil is extracted and inserted in a separate image, in which all the pixels that are not part of it are placed at NaN. The reason for this choice is that for the resize of the pupil there are two cases: *(i)* the pupil grows downwards. In this case the resize must be carried out vertically, and *(ii)* the case in which the pupil grows to the right or left, requiring the resize to be done horizontally. To ensure that the growth is homogeneous along the two axes, however, the pupil must be exactly in the center of the image, otherwise, the growth would be disproportionate, with a larger increase towards the direction in which the center of the pupil relative to the center of the image. Then, the image containing only the pupil is cropped so that it is centered, and the various dimensions for the original image are maintained. In this way, when the new pupil is overwritten on the starting image, each pixel goes into place correct and the mapping between the two images is respected. The image containing the pupil is enlarged using a function that takes as input a resize parameter "nearest", which ensures a nearest-neighbor interpolation: each output pixel of the resized image has assigned the closest

pixel value of the original image. In practice, this serves to give the pupil, after resizing, a more rounded shape, to reduce the loss of image quality. The pupil thus obtained must now be overwritten on the original image. First, a check is made on the dimensions of the axes of the ellipse that describes the new pupil, to prevent it from growing outside the contours of the iris. These dimensions are then used to automatically trace a new truth mask, specific for the current image, which will then be used to evaluate the precision of the segmentation algorithm.

- **Overwriting of the Pupil** At this point, the pupil, which has been enlarged horizontally or vertically, depending on the direction on which it must expand, is overwritten on the original image only for the half of interest, one pixel at a time, up to the last pixel of each. row/column, which will be the new pupil limit for that row/column. Starting from the original limit, a one-dimensional segment is extracted, which reaches the limit of the iris on the same row/column, depending on the current direction. This segment is resized, so that its size decreases according to the new pupil limit, and overwritten in the new position it must occupy. Figure 4.18 shows an example of resizing the pupil downwards, using a scaling value of 1.5, with the consequent compression of the iris in the area where the pupil expands, to avoid losing its distinctive characteristics.

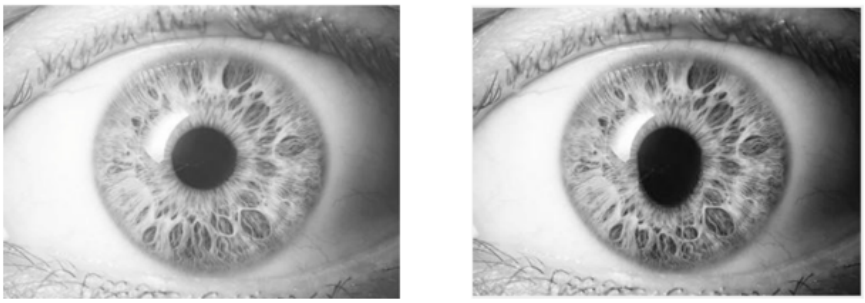


Figure 4.18: Resize down with scaling 1.5

When the pupil is enlarged homogeneously, as before, a key role is played by the pixels that make up the outer contour of the pupil, which in this case undergo a diagonal translation and no longer only horizontally or vertically. Figure 4.19 shows the process that leads, starting from the original image, to the pupil obtained by resizing for a certain scaling value. In the fourth image, the two concentric circles, added for demonstration purposes, represent the boundaries of the pupil before and after scaling. To map each pixel of the new

pupil to the initial one, its original coordinates are obtained using the following equation:

$$xOriginal = \frac{widthOriginal * x}{newWidth} \tag{4.4}$$

$$yOriginal = \frac{heightOriginal * y}{newHeight} \tag{4.5}$$

where x and y represent the coordinates of the current pixel, of which we want to find the original coordinates ($xOriginal$ and $yOriginal$) in the starting image, whose dimensions are $widthOriginal$ and $heightOriginal$, while $newWidth$ and $newHeight$ those of the image containing the isolated pupil, zoomed by a certain scaleOverall value. At this point, the segments of the iris adjacent to that pixel in the original image are extracted, resized, horizontally or vertically, and overwritten in the new position. In addition to overwriting to resize in a single direction, a necessary operation concerns the resize of the four quadrants of the iris, seen in the subsection **??**, which are not adjacent to the pupil, and therefore with the operations previously described they do not undergo variations, and are overwritten in some places. To overcome the problem, a double iteration is carried out on the image: one on the rows, in which each horizontal segment belonging to these quadrants is narrowed in width, the other on the columns, in which each vertical segment is narrowed in height. At the end of this operation, each pixel will have undergone a diagonal translation, based on the extent of scaling on the pupil.

The proposed extension

The extension of the Daugman algorithm has been specifically designed for the detection of the characteristics of eyes affected by pupil malformations, due to Coloboma. As seen in the previous section, for which the Daugman algorithm does not guarantee efficient results, as there is the need to go beyond the original assumption for which the iris and pupil can be described by two internal circumferences. This extension is mainly focused on the first segmentation phase, in which the pupil is not identified through a circumference but through an elliptical shape, based on the implementation of Libor Masek [153]. A first mention of the possibility of recognizing the pupil through an ellipse, and specifically through a point-by-point segmentation, was made by Daugman in the publication "New Methods in Iris Recognition" [57]. In this paper, he underlined the limitation concerning the capture of the iris for which there is the need to align the eye to the optical axis of the camera. In the following I describe the modification addressed to the phases of the Daugman's algorithm to detect puplis affected by Coloboma.

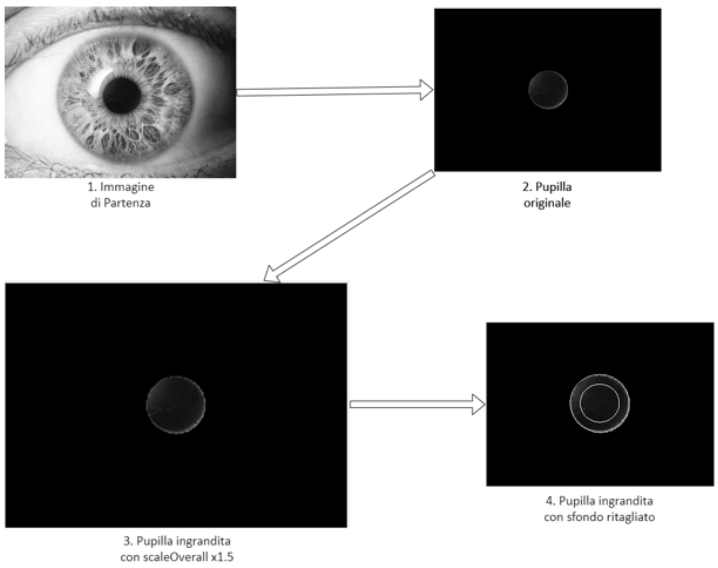


Figure 4.19: Overall pupil scaling process

- *Segmentation.* Since the algorithm is designed for cases in which the pupil undergoes deformation due to the Coloboma, the assumption that the eye is identifiable through two internal circumferences is relaxed. Therefore, a first external circumference is searched because the iris does not undergo any deformation in its outline. Moreover, the approximation of an ellipse that represents the pupil is searched, since it hardly has a regular elliptical shape. Figure 4.20 shows, at a high level, the main phases of the segmentation process, with the first step of localization of the iris, followed by the localization of the pupil through an ellipse which is finally subjected to an approximation to the circumference.

  A pre-processing phase is carried out to identify the contour of the iris, including noise reduction, contrast increase, and filling of the iris area. A Gaussian smoothing filter is used for noise removal, keeping the edges intact, followed by an increase in contrast. A filling operator is applied to the result to make the area inside the iris uniform, followed by a further increase in contrast and sharpness.

  The fundamental points of the algorithm, on which much of the final result was concerned, were mainly identified in three parts.

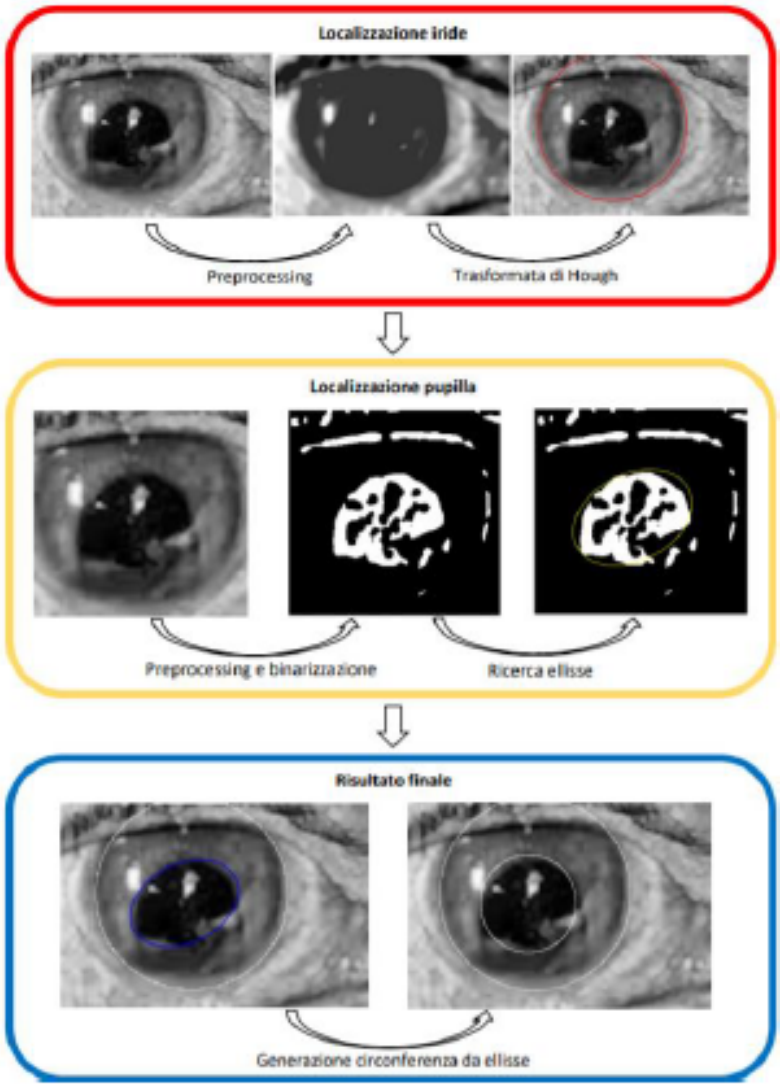


Figure 4.20: Complete Segmentation Process.

1. *Sensitivity for the recognition of the iris*: For the detection of the iris, the Matlab function $imfindcircles()$ is used, which identifies circles inside

the image, according to the input parameters. The algorithm chooses based on the dimensions of the rays, in particular chooses the one with minor radius, to prevent the sclera from being segmented in place of the iris. One of the input parameters is represented by the sensitivity for the accumulator array of the circular Hough transform, a standard computer vision algorithm that can be used to determine the parameters of simple geometric objects present in an image, resistant to noise and partial occlusions. The "ObjectPolarity" parameter is set to "dark", as the circular contour sought has a darker shade of gray than the outside, the iris being darker than the sclera, tending to white. Another important parameter is the sensitivity in the search for circumferences, with a range of values included in the interval [0,1], with a default value of 0.85, we tried to increase this value set to 0.93. Increasing this parameter leads to the detection of a greater number of circular objects in the image, including some that are more obscure, but at the same time increases the rate of false positives. In some cases, especially when the pupil began to assume relatively large dimensions, tending to those of the iris, there were situations in which the iris was detected outside its real borders, consequently leading the iris to be recognized as a pupil. Therefore, several attempts were made, trying to find the optimal value, which in the end was identified as 0.85, which is the default value itself. The tests that led to trying values below this threshold did not lead to improvements, and on the contrary, they turned into a greater difficulty for the algorithm, in certain images, to detect the circles that could represent the iris, returning situations error in segmentation. Following the identification of the circumferences, the one with the smaller radius is chosen because, after the application of the $imfill()$ function, the internal region of the iris may present in addition to the pupil other circumferences due to light reflections or other factors external. They are in any case ignored since they do not reach the minimum size threshold necessary to be taken into account by the algorithm. Figure 4.21 shows a low-level representation of the described process of localization of the iris, with the main functions used, and specifically the modifications to which the original image is subjected for correct final identification of the contour of the iris.

2. *Sensitivity for pupil recognition*: After identifying the outline of the iris, the pupil is localized within the circumscribed area in the original image. Also, in this case, a preprocessing phase is carried out to reduce noise and increase contrast. Considering that the pupil is generally the darkest area of the eye, binarization is used to allow easy extraction of its region, using the $imbinarize()$ function, to facilitate the detection operation of the ellipse, to be carried out through $regionprops()$. To emphasize the

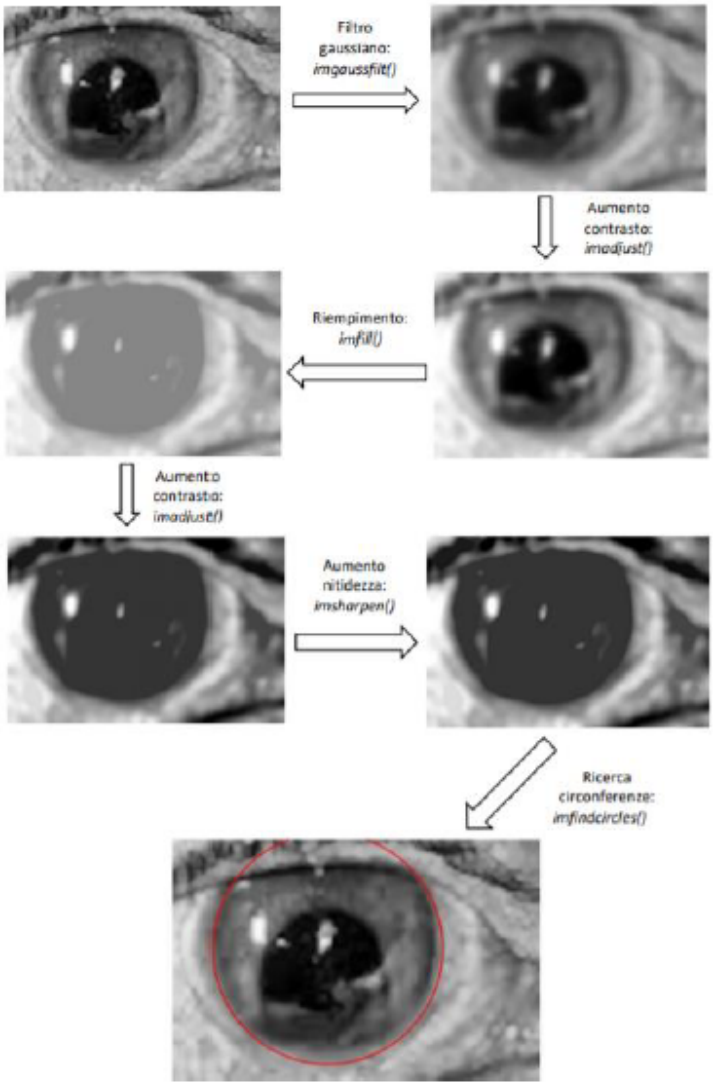


Figure 4.21: Localization of the iris.

darkness of the pupil "$ForegroundPolarity$" parameter is set to "dark"; the binarization function also requires a sensitivity parameter, whose de-

fault value is 0.5 and is originally set to 0.25. In this case, the increase in sensitivity leads to include, within the threshold, more foreground pixels, with the risk of including some unwanted backgrounds. Since the original value is already relatively low, compared to the default, this problem has practically never occurred. On the contrary, however, we found some cases in which a region inside the pupil was detected as a pupil, perhaps because of a reflection of light, and which therefore led to an incorrect, or only partial, detection.

3. *Check the pupil area*: For the identification of elliptical regions, the function $regionprops()$ is used, which identifies circular or elliptical shapes in the image, measuring and returning their characteristic properties. For the regions thus identified, selection criteria have been established to take into consideration only the most relevant: first of all, those with an area that is too large, greater than 1/3 of the area of the image, or too small, with more than 20% of the perimeter points outside the image. At this point, of the remaining ellipses, the one with the lowest average pixel intensity is chosen, since the pupil is the darkest region of the eye Starting from the elliptical contour identified, a circumference is constructed having the mean of the two axes as a radius, to allow the translation of Cartesian coordinates into polar coordinates, as described in the normalization phase of the Daugman algorithm [58]. Figure 4.22 shows a low-level representation of the pupil identification process, initially through an ellipse that is approximated in circumference, again exposing the changes to which the original image is subjected for segmentation.

   But we noticed that this represented a problem in images of this type. Since the iris represents a large part of the image, as it grew, the pupil came to exceed the predetermined threshold, thus not being detected by the algorithm. In the case of images captured in this way, therefore, the algorithm would have had problems in detecting the pupil, and for this reason, it was considered appropriate to remove this control and replace it with the verification that the pupil area should not exceed that of the iris, which allowed the detection, albeit not perfect.

The identification of the contours of the eyelids is carried out through the Radon transform, taking into consideration the regions of the image ranging, respectively, from the upper limit of the pupil to that of the iris, for the upper eyelid, and from the lower limit of the pupil. at the lower limit of the iris, for the lower eyelid. An edge map is then generated, on which the Radon transform is applied, and among the segments identified that exceed a certain threshold, the one with the maximum value in the parameters is chosen and all the pix-

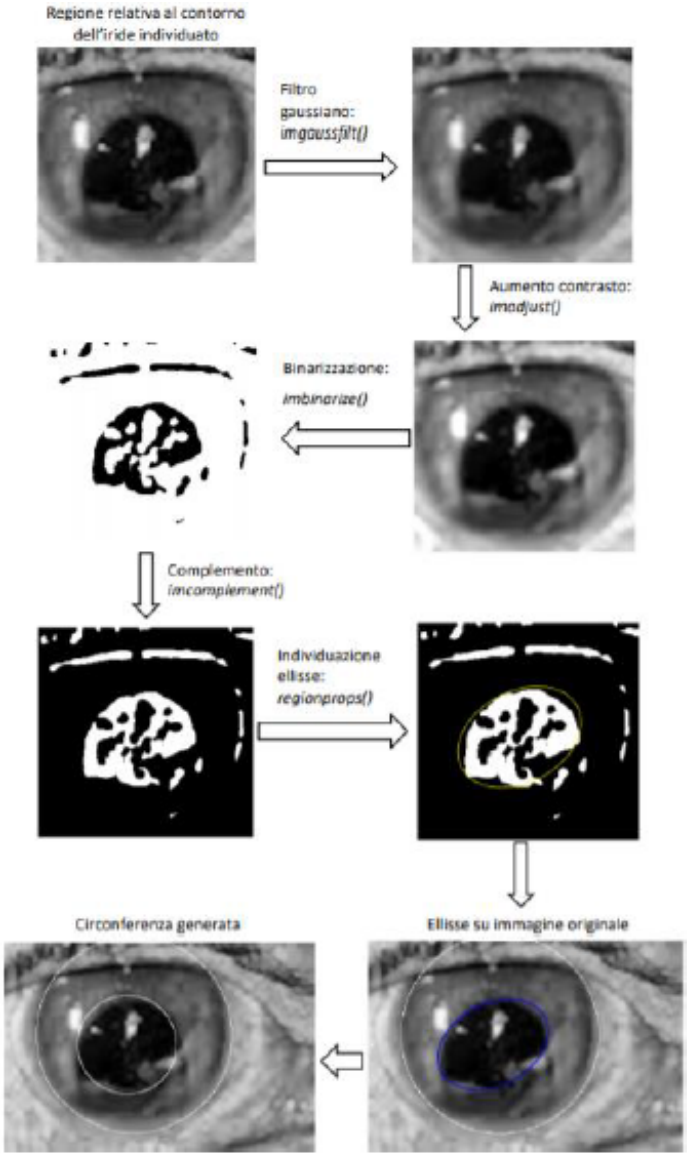


Figure 4.22: Localization of the Pupil.

els belonging to the corresponding region are placed at NaN. Instead, for the removal of lashes and reflections, two thresholds are simply used: pixels with a value lower than 30 are considered lashes, while with a value greater than 240 reflections, and in both cases placed at NaN. Figure 4.23 shows in detail the steps that return as a result the final image, deprived of the noise zones. these zones will then be fundamental in the encoding phase, since it will be taken into consideration for the creation of a noise mask used in the matching between irises.

- *Normalization* The implementation of the normalization phase is based on Daugman's rubber-sheet model: a certain number of radial lines are drawn from the center of the pupil, which defines the angular resolution, and on them some points, which instead define the radial resolution. Given the non-concentricity of the circumferences of the iris and pupil, the following remapping technique is adopted:

$$r^{'} = \sqrt{\alpha}\beta \pm \sqrt{\alpha\beta^2 - \alpha - r_I^2} \tag{4.6}$$

$$\alpha = o_x^2 + o_y^2 \tag{4.7}$$

$$\beta = cos(\pi - arctan(\frac{o_x}{o_y}) - \theta) \tag{4.8}$$

where o_x and o_y respectively represent the horizontal and vertical displacement of the center of the pupil with respect to the iris, $r^{'}$ is the distance between the contour of the pupil and that of the iris, for an angle θ, while r_I is the radius of the circumference describing the outline of the iris. With normalization, a two-dimensional array containing the extracted pixels is obtained, having a width and height respectively equal to the angular and radial resolutions. A second array of the same size enables us to mask the noise zones detected during the segmentation phase. Figure 4.24 shows the complete normalization phase, applied to Figure 4.23, and therefore already subjected to the entire segmentation process, for iris and pupil, and noise removal. At this point, the masks are generated starting from the normalization.

- *Coding.* Starting from the two-dimensional normalized pattern the coding phase, for each row, corresponding to a circular ring of the iris region, a convolution was carried out by means of the one-dimensional Log-Gabor Wavelets. The result is a bit template, and a corresponding noise mask, whose total number of bits is given by:

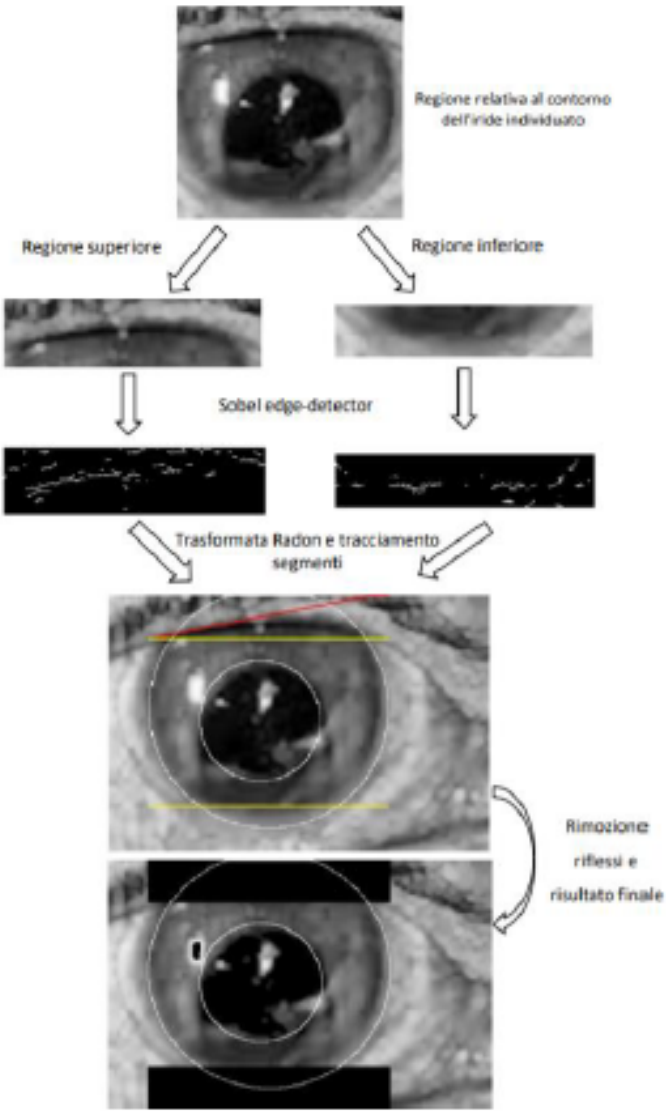


Figure 4.23: Localization of the Pupil.

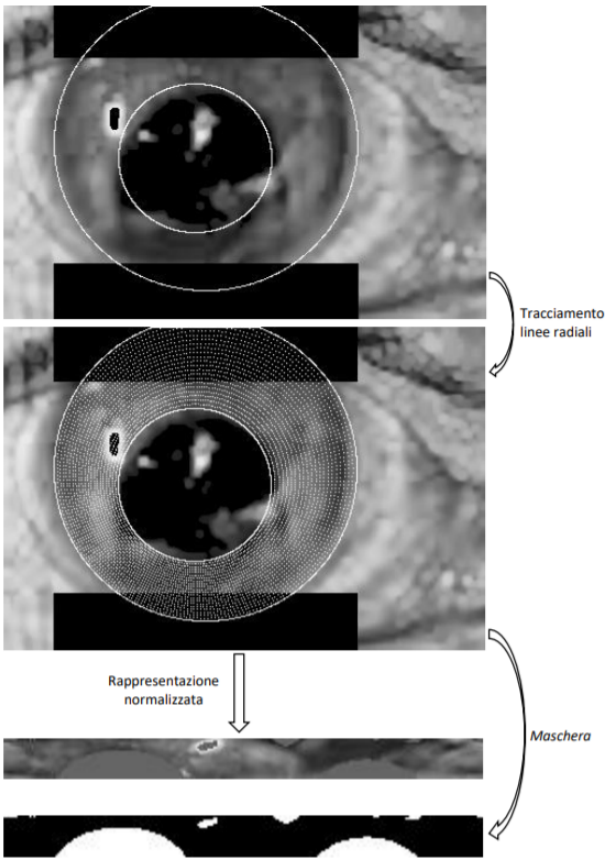


Figure 4.24: Normalization process.

$$N = 2 * angularresolution * radialresolution * numberof filters \quad (4.9)$$

which in the specific implementation corresponds to:

$$N = 2 * 20 * 240 * 1 = 9600bits \quad (4.10)$$

Figure 4.25 shows, in order, the results obtained through the implementation of Libor Masek, the implementation of the circle and ellipse and the comparison between the errors, combined, of the two implementations. For the Libor

Masek algorithm, the cases in which the segmentation has failed, returning an error, have neither been included in the graph nor taken into account in the calculation of the average error. For the implementation of the circle and ellipse, on the other hand, in case of failure in the detection of the pupil, the error was set to 1. Finally, in the comparison graph (Figure 4.26 and Figure 4.27) of the errors of the two implementations, an error in the segmentation, both for pupil and iris, was considered by setting the distance to 1.

The implementation of circle and ellipse achieved much better average results, with an average improvement of 2.75 in the detection of the iris, 1.56 for the pupil and 1.8 overall. In both cases, there were more difficulties in segmenting the pupil, due to its irregular shape, as well as the noise present in some images. Figure 4.28shows an example of the detection process that is applied to each image of the dataset, with the expected segmentation on the left, calculated during the Data Augmentation phase, and on the right the segmentation detected by the algorithm.

We analyze the results of the iris and pupil caption, shown in Table 4.2 and Table 4.3, respectively. For the calculation of the overall results, as shown in Table4.4 for each iteration, the sum of the distances of the iris and pupil were considered for the respective image, therefore the values shown are to be considered on a maximum of 2.

Table 4.2: Iris caption

image	average	sd	max
1	0.043	0.045	0.164
2	0.011	0.006	0.036
3	0.077	0.014	0.11
4	0.017	0.014	0.092
5	0.034	0.010	0.054

- *Matching* The matching phase computes the Hamming distance, defined as the sum of the disagree bits (sum of the exclusive OR between X and Y) over N, the total number of disagree bits in the template for the two irises is calculated, with the respective noise masks, using the formula:

  Each region of the iris will produce a bit pattern independent of that produced

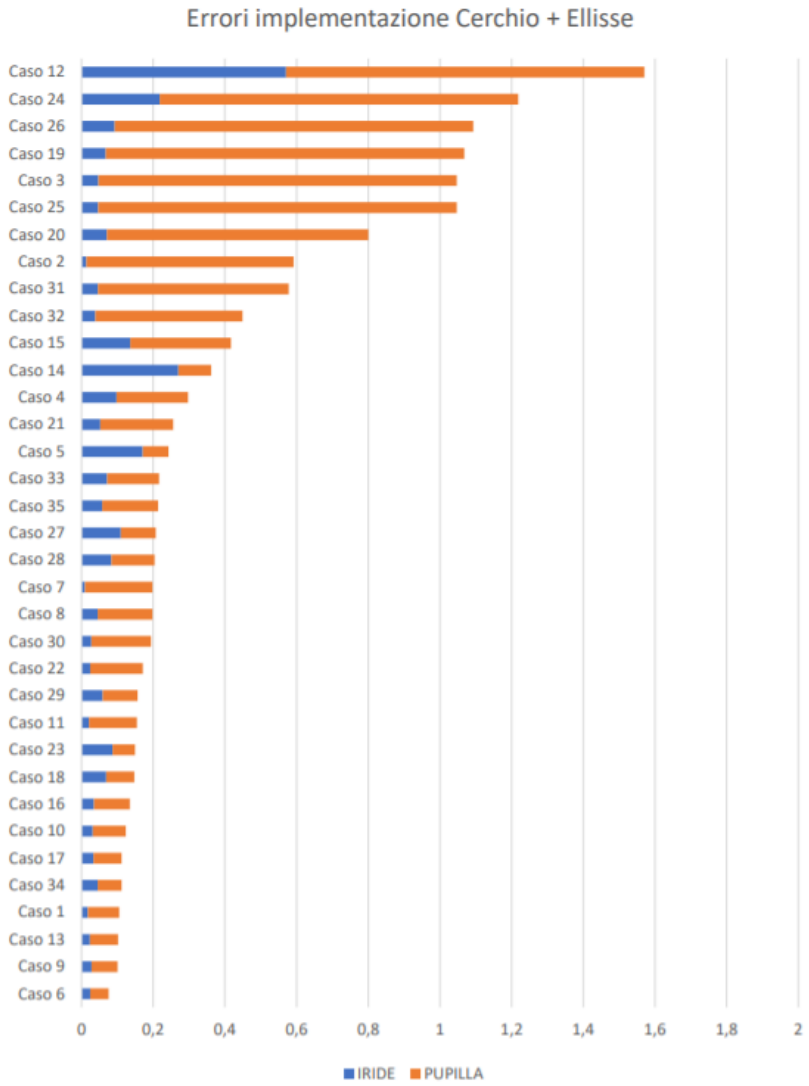


Figure 4.25: Errors in the implementation of circle and ellipse.

by another iris. If two-bit patterns are completely independent, such as templates generated by different irises, the Hamming distance between the two patterns should be 0.5. In the calculation of the Hamming distance, the bi-

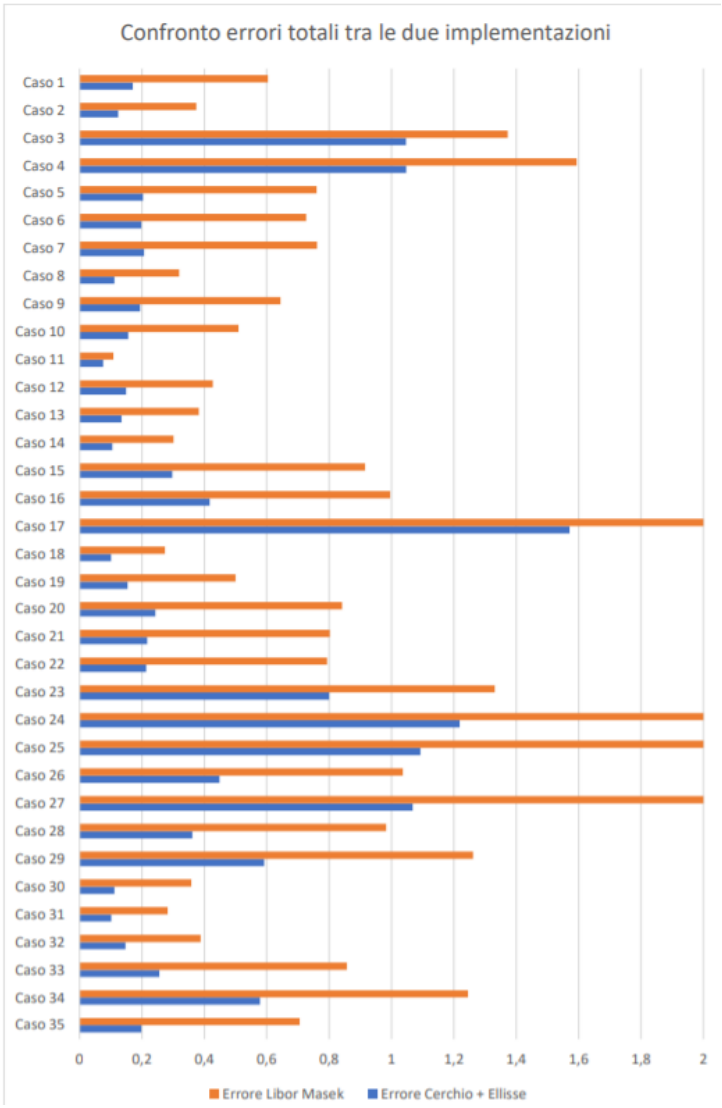


Figure 4.26: Comparison of the overall errors of the two implementations.

nary masks for noise removal are also considered: given two templates $codeA$ and $codeB$ and the corresponding masks $maskA$ and $maskB$, the Hamming distance is calculated considering only the significant bits for both templates,

	IRIDE	PUPILLA	TOTALE
LIBOR MASEK	0.22	0.5	0.72
CERCHIO + ELLISSE	0.08	0.32	0.40

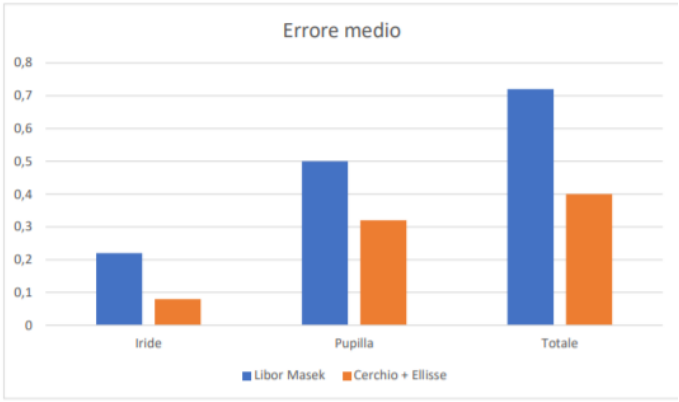


Figure 4.27: Comparison between the average errors in the two implementations.

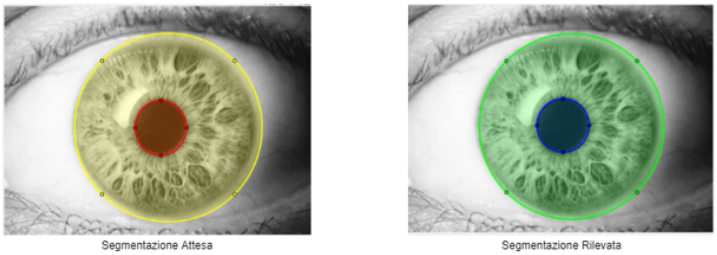


Figure 4.28: Detection example

i.e. the bits of the templates whose corresponding bits in the masks are set to 1. To overcome the problem of the rotation of the iris in the images, one of the two templates is shifted by a few bits first to the left and then to the right, for a certain number of times, calculating the Hamming distance for each shift and

Table 4.3: Pupil caption

image	average	sd	max
1	0.066	0.032	0.122
2	0.085	0.047	0.172
3	0.085	0.034	0.162
4	0.048	0.023	0.117
5	0.034	0.017	0.073

Table 4.4: Overall surveys

image	average	sd	max
1	0.11	0.067	0.262
2	0.096	0.051	0.196
3	0.163	0.035	0.231
4	0.066	0.024	0.142
5	0.069	0.018	0.113

$$HD = \frac{\|(codeA \otimes codeB) \cap maskA \cap maskB\|}{\|maskA \cap maskB\|}$$

Figure 4.29: Formula to find the error

considering the lowest one as the best, as it corresponds to the best match between the two templates. In this phase, the inconsistency in the rotation of the

images is also taken into consideration, and to do this, for each pair of images compared, one template is shifted concerning the other, and the Hamming distance is calculated for each shift. Finally, the smaller Hamming distance among those calculated on all the shifts performed on the templates is taken into account. While comparing two images of the same iris, due to numerous external factors, it is very rare to obtain exactly 0.0 as a result. For this reason, the result obtained from the matching must be analyzed and a threshold must be established, through the observation of intraclass and interclass comparisons, with the related False Acceptance and False Rejection Rate.

To generate an iris identification code, a variant of the algorithm set out in the publication "Iris Code Generation and Recognition" [93] was used, consisting of 4 steps:

1 Starting from the iris template, with a size of 64x512, it is divided into 16 blocks of 64x32 pixels;

2 Each of these 16 blocks must be converted into one bit, for a total of 16 bits, using the average of the pixels of the entire template and each block;

3 For each block, if its average is higher than that of the template, a value of 1 is assigned;

4 Otherwise, the value 0 is assigned.

The code obtained in this way can be used for matching with the coding of another iris obtained in the same way, through the Hamming distance.

This algorithm has been adapted in such a way as to make the most of its functionality. First of all, the generated iris template has dimensions of 20x480 pixels, and in addition to it, a second template of the same size is also generated, representing the noise mask present in the original image. Both of these templates are then converted into the corresponding code, applying the described algorithm, and the Hamming distance is also calculated to also exploit the information relating to the noise in the image. The matching operation is then carried out iteratively. For each subset of images, we start from the original, with overall scaling equal to 1.0, which is compared with all images with equal overall scaling and gradually increasing scaling in the 3 directions. Once the image with subsequent overall scaling is reached, i.e. in the first case 1.1, it is compared for one last time with the previous scaling image (1.0), and then becomes the image with which the subsequent comparisons, up to the next overall scaling.

This process, described graphically by Figure 4.30 and Figure 4.31, aims to provide greater consistency and coherence to the matching between the obtained encodings: the fact of comparing each modified image with the starting point,

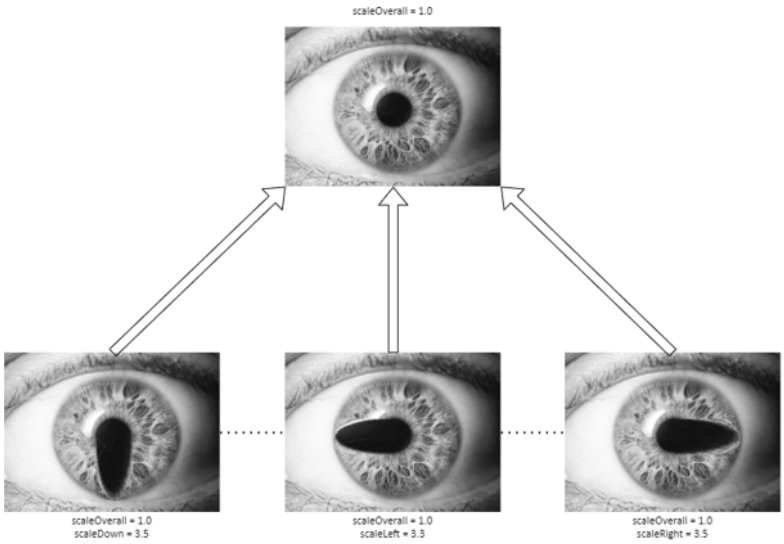


Figure 4.30: Representation of the iterative process of matching between images with the same scaleOverall.

in which the pupil still had a circular shape, allows you to simulate a possible real case in which the pathology occurs, with the physical consequences that it brings, and to evaluate what level of reliability the algorithm guarantees us.

The output of a matching carried out on the template of two different images outputs a value within the interval [0.1], and the lower it is, the greater the correlation between the two irises. Since the template consists of a matrix of binary values, in the case of matching between two images of different eyes, since the templates of the irises are completely independent of each other, each bit has a 50% probability of being 1 and 50% to be 0, so the expected result is about 0.5. Furthermore, in the matching phase shifts are carried out on the template to take into account the rotation of the image; for each of them, a Hamming distance is calculated, among which the minimum value is chosen, the average value in comparison between different classes it will be slightly below 0.5. When the two images instead depict the same eye, a lower value is expected, but due to various external factors, such as pupil dilation, image quality, noise sources such as light or eyelashes, distance or l skew of the capture, it is very difficult to get an output of exactly 0. Therefore, the goal in the context of the matching phase is to observe a large number of outputs, both for intraclass and interclass comparisons, and to establish a threshold within

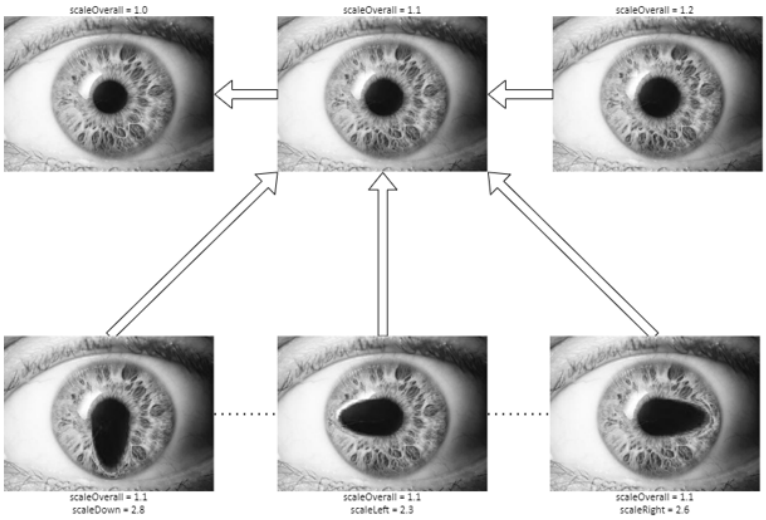


Figure 4.31: Representation of the iterative matching process at the change of scaleOverall.

which it is still possible to estimate a reliable correlation between two irises. A discussion on the threshold settings is reported in Appendix A.

Preliminary results

This is an ongoing work. Thus, I report the results of a preliminary evaluation obtained by applying on the synthetic dataset the iterative matching process, described in subsection 4.2.4. In particular, I analize the number of times that detection of the Hamming distance exceeds the predetermined acceptance threshold. Because we consider the context of a series of intraclass we obtain an estimate on the False Acceptance Rate of the analyzed data, reported in Table 4.5.

For a more detailed analysis, in addition to conduct the analysis on the whole dataset, I also consider the data relating to the growth of the scaling of the images taken into consideration. In this way, it is possible to get a clearer idea of how the results of the biometrics process vary as the pupil size increases. Each scale value, represents the scaling limit taken into consideration at each iteration and is increased each time at intervals of 1.5, up to 3.5, i.e. the maximum scaling value at which in most cases the pupil grows, before reaching the limits of the iris.

The slight increase identified for the scaling levels 2.5 and 3 is mainly due to the fact that some images have a more dilated pupil already at the start. For this reason,

Table 4.5: Results obtained for the entire Dataset

Scale value	Excess	False Acceptance
1.5	61/1068	5.71%
2	138/1772	7.79%
2.5	175/2147	8.15%
3	189/2323	8.13%
3.5	190/2381	7.98%
Tot	191/2387	8.00%

they are subjected to a smaller number of scaling degrees, and therefore these represent the maximum scaling value they can get before reaching the limits of the iris. The whole False Acceptance error is 8%.

Chapter 5

A.I. IN ONCOLOGICAL DISEASES

This chapter aims at answering the following research question:

RQ2: How Artificial Intelligence may support the oncological disease classification?

To answer this question we focused our attention on two disease: Leukemia and Melanoma.

For malanoma desease we propose (C03 - C04) a methodology for the classification of melanoma by adopting different deep learning techniques applied to a common image dataset composed of images from the ISIC dataset and consisting of different types of skin diseases, including melanoma on which we applied a specific pre-processing phase. Our aim is to compare the results of the adopted techniques in order to select the best effective neural network for the recognition and classification of melanoma, and evaluate the impact of the pre-processing phase on the final classification. We have also developed an application for the recognition of the lesion based on augmented reality as a support tool for clinicians in the diagnosis of the melanoma (J03). As for leukemia, however, we define a process aiming at detecting a set of differentially expressed genes in terms of methylation level, i.e., genes that in different conditions have an expression level significantly different in the AML and ALL cases, and their characteristic pathways. The detection of gene expression data samples involves feature selection and classification (J04). To this aim, we adopt Deep Learning models (e.g., feature selection techniques and classifiers methods). The analysis has been performed on a dataset consisting of samples from people with leukemia, characterized by a fixed list of genes; the samples belong to two distinct classes: ALL and AML.

5.1 Supporting Melanoma detection

In this section, after a brief introduction on the main concepts concerning Melanoma, I present the results related to the Melanoma detection and the supporting real-time application we developed for assisting the clinician during the diagnosis.

5.1.1 Melanoma

Melanoma is a tumor, often very aggressive, which originates in the skin or, more rarely, in the eyes or mucous membranes [188]. It is originated when a genetic error occurs in melanocytes, located in the basal part of the epidermis. Melanocytes are cells located in the lower part of the epidermis, just above the dermis, producing a pigment called melanin, which gives colour to the skin.

Melanoma is a neoplasm visible to the naked eye and originates from a pre-existing mole that changes shape or colour or from the appearance of a new mole on intact skin. It is possible to distinguish various melanoma clinical development stages (i.e., clinical classification) that are associated with a different morphological aspect, and also some different histological families (i.e., histological classification). The behaviour of the neoplasm is influenced both by the clinical and the histological aspects. Melanoma is an always a malignant neoplasm. Indeed, it is never possible to define a benign melanoma, at most we can speak of a benign neo that does not have the characteristics of melanoma. skin followed by dermoscopic analysis, biopsy and histopathological examination [74].

The incidence rate of skin cancer has increased over recent years [123]. Among skin cancers, melanoma is the most frequent and deadliest. Cutaneous melanoma will be created by abnormal growth of melanocytes [150] and this cancer is primarily caused due to severe exposure of skin to the sunlight. It is responsible only for 4% of all cancers occurred in human skin, but it is accountable for 75% of deaths caused by skin cancers [143]. If the cancer is detected preventively and treated properly, the chances of treatment increase [13, 84]. The American Cancer Society estimates that in 2018, about 91.270 new melanomas are suspected to be diagnosed (about 55.150 in male and 36.120 in female) in the United States. Almost 9.320 people are presumed to die of melanoma (about 5.990 male and 3.330 female)[1]. Australia and New Zealand have the highest rates of melanoma in the world [27].

The manual inspection from dermoscopy images made by dermatologists is usually time-consuming, may be subject to errors and it is a subjective analysis. However innovative early detection programs, in combination with improved diagnostic tools and new immunologic and molecular target treatments for advanced stages of the disease, can improve the treatment of this disease [226].

[1] https://www.cancer.org/cancer/melanoma-skin-cancer.html

Automatic recognition approaches are highly demanded, but they are not a very easy task. Distinguish melanoma from non-melanoma figures is very hard. The approaches to the automatic recognition of melanoma are in great demand, in fact, the recent development of technologies and models of machine learning has led to an expansion in the use of these techniques in the biomedical field. A system for the computerized diagnosis of a skin lesion, provides a quantitative and objective assessment of the skin lesion, as opposed to visual assessment, which is of a subjective nature. This process consists of four main phases: image acquisition of the skin lesion, segmentation of the lesion, extraction of the characteristics, and classification of the lesions [242].

Typically these lesions are pigmented, with a diameter smaller than 6 mm, flat or slightly raised and of a uniform color. The nevus constitutes the benign form of the tumor and is characterized by fusiform cells that can evolve into i) junctional nevi, which accumulate at the junction between the dermis and epidermis, ii) compound nevi, which accumulate both in the dermis and in the epidermis, and iii) intradermal nevi, which result from a migration of junctional nevi into the underlying dermis. A melanoma may be in various clinical stages of development (clinical classification) that are associated with a different morphological aspect. The behavior of the neoplasm is influenced by both the clinical and histological aspects [175]. During a visit, a melanoma can appear at the inspection of four types:

1. *Flat non-palpable melanoma*: it represents the most frequent form (70%); tends to grow towards the outside rather than the inside;

2. *Cupoliform or Nodular Melanoma*: it is a variant of melanoma with rapid evolution and with a high risk of progression which tends to appear at an older age. It represents 10-15% of all melanomas.

3. *Lentigo Maligna (melanoma in situ)*: it is a slowly evolving lesion that manifests itself as a flat, non-palpable, brown, very smooth spot, with loss of the normal skin profile. It generally has a slow growth rate (years) and rarely spreads to other parts of the body.

4. *Acral-freckled Melanoma*: it appears instead in the acral areas (palm of the hand, sole of the foot) represents 5% of all melanomas.

Due to the extreme heterogeneity of the lesions, it is very difficult to identify and diagnose melanoma early and the importance of diagnosing melanoma early is not to be underestimated, this is because the prognosis in melanoma is directly proportional to the depth of the neoplasm [192].

The conception of the melanoma ABCDE mnemonic was a milestone for the early detection of melanoma. Before the 1985, the diagnosis of melanoma was often made considering macroscopic features such as bleeding and ulceration of the

lesion, but since these features are often found in advanced lesions, they were considered of little use in recognition early melanomas [192].

The ABCD rules were proposed by Friedman *et al.* in 1985 [86] to create a direct and simple interpretation tool for clinicians. Subsequently, Abbasi *et al.* [1] added the letter E (for evolution) as a recognition criterion as a rapidly evolving lesion can be a melanoma. The effectiveness of the ABCDE system has been validated in numerous studies conducted by dermatologists [34, 48]. "ABCDE Rule" (A = asymmetry, B = edges, C = color, D = size, E = evolution) [85]. The $asymmetry$ in the form is a very important element in the diagnosis of melanoma. Benign neo is generally circular or rounded while melanoma is more irregular and can be flat or raised. As for the $edges$, neo benign has more clear edges and instead melanoma has irregular and indistinct edges in most cases. The $colour$ in melanoma is often variable and it can be brown, black, reddish and change over time. Furthermore, the $size$ of the nevi is suspicion when the diameter is greater than 5-6 millimeters. Finally, it is important to consider the $evolution$ of a nevus, because a melanoma in a rather short time tends to grow quickly.

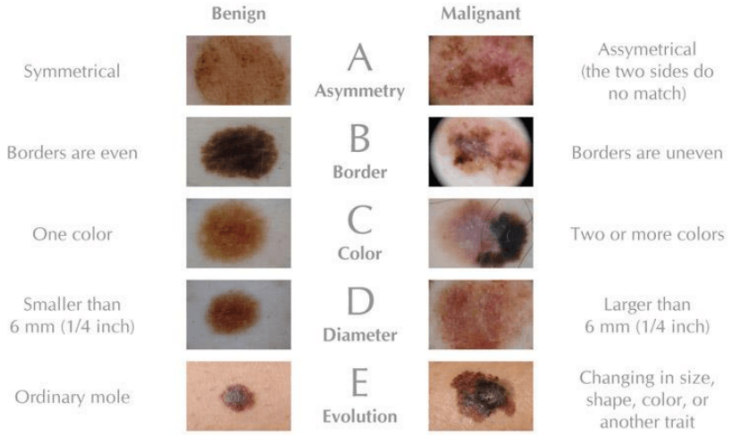


Figure 5.1: The ABCDE rules.

5.1.2 Dataset

In this work, the dataset used to train and evaluate the Deep Neural Networks is HAM10000 ("Human Against Machine with 10.000 training images") [221], containing images of the ISIC dataset. The dataset is composed of 10.015 dermatoscopic images divided into 7 target classes, as Actinic keratoses and intraepithelial carci-

noma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis like lesions (solarlentigines/seborrheic keratoses and lichen planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and haemorrhage, vasc). The predominant part of the dataset is composed of images of melanocytic nevi (nv 6.705), benign lesions ilar to keratosis (bkl 1.099) and melanomas (mel 1.113).

5.1.3 A Comparison of Neural Network Approaches for Melanoma Classification

Before starting with the analysis process, it was necessary to make an accurate study on the structure of the dataset. To this aim, *i)* we binarized the problem by working on the appropriate labels of the above-mentioned dataset, *ii)* we organized the images, *iii)* we defined some operations to perform an image processing on the images present in the dataset. Thus, we first transformed the problem into a binary classification problem. Given the large number of classes, we have binarized the problem, which has made all the labels other than *"mel"*, equal to *"not mel"*, in so that you can clearly distinguish which images portray a melanoma and which are not. This transformation returned an unbalanced dataset where the distribution of the melanoma class is skewed in contrast to non-melanoma, as shown in Fig. **??**. To mitigate this problem, after the phase of image processing has been carried out we defined and implemented a Data Augmentation process.

The organization of the images was done dividing, thanks to the control on the labels, the images related to melanoma and not melanoma in two separate folders. Finally, about 10% of the images has been pruned because they had occlusions that did not allow a good segmentation of the skin lesion.

The key steps in a computer-vision based diagnosis of melanoma classification are: image acquisition, pre-processing, segmentation, extraction of features and classification [214].

Pre-processing

Concerning the pre-processing, there are three main steps to automatically extract a lesion in a dermoscopic image. These steps strongly depend on the clinical features of a lesion. The operations that have been performed are:

- *Hair removal.* Hair occlusion in dermoscopy images affects the diagnostic operation of the skin lesion. We have used *Canny edge detection* for the hair removal, that includes two main stages: in the first step, light and dark hair is segmented through adaptive canny edge detector and refinement by morphological operators. Finally, in the second step, the hairs are repaired based on

the painting of the transport of coherence to several resolutions [219]. Following the Canny edge detection, the *Otsu's threshold* has been used to obtain a black and white image to be used as a mask for hair removal. Once the mask is obtained through the Otsu's threshold, the dilation operator has been used to grant greater precision in hair capture. Finally, we performed image inpainting, a technique that performs a sort of interpolation for digital image processing to reconstruct parts of damaged digital images [51]. The hair removal process is shown in Fig. 5.2:

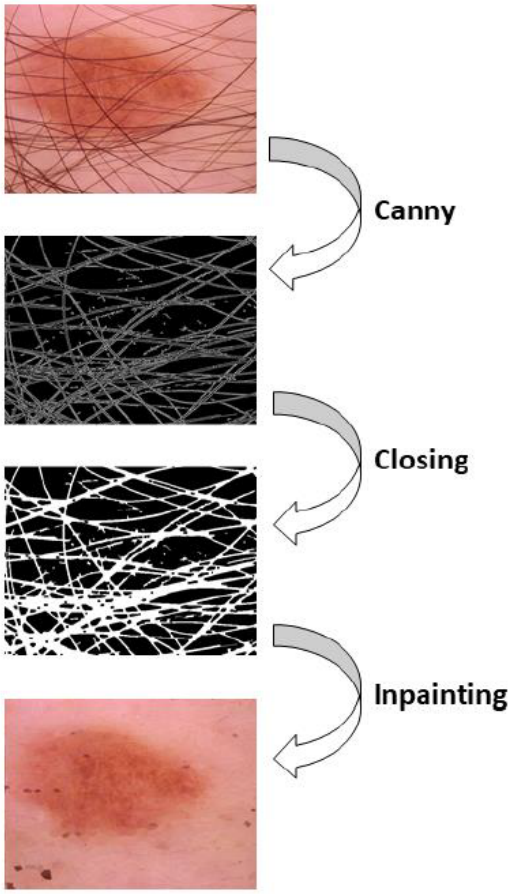


Figure 5.2: Hair removal by using the Canny edge detector.

- *Lesion segmentation.* Segmentation is one of the most important techniques

that lead to the processing of data from images. Its primary purpose is to select specific objects or regions in an image. The image is divided into homogeneous regions concerning a choice of properties, such as brightness, colour, texture, etc. The success of the subsequent phases depends exclusively on a good segmentation capable of extracting the regions of interest (ROI) of the dermoscopic image. At this stage we transforms the image, from RGB to grayscale and the lesion is separated from its background, i.e., the skin, this is done by binarizing the image to identify the background of the skin to be discarded and the region of the lesion to be analyzed. Subsequently, the edges are extracted to obtain information on the symmetry and regularity of the edge of the lesion [168].

For the segmentation of the images we used Otsu's Thresholding. This method identifies which are the relates of a lesion and is based on two threshold values, i.e., a minimum and a maximum value.

The Otsu's Thresholding method is used to automatically perform clustering-based image thresholding or, the reduction of a gray level image to a binary image. The algorithm assumes that the image contains two classes of pixels following bimodal histogram (i.e., foreground pixels and background pixels), it then calculates the optimum threshold separating the two classes so that they are combined spread (intra-class variance) is minimal [223]. It is assumed that the edges with a gradient of intensity greater than the maximum value are edges, while those below the minimum value are certainly not edges and therefore to be discarded. The process is shown in topmost part of Fig. 5.3.

- *Clinical feature segmentation.* In this case there is a local segmentation of the lesion, thus highlighting the clinical features of a lesion such as the texture, shape and color. Subsequently, it was necessary to reduce the over-segmentation, a typical "problem" that occurs in the output of the previous phase. by exploiting a technique based on temporal and spatial averaging. In this case, a median filtering was used which preserves the edges by removing the over-segmentation and also the blurring effect. The median filter consists of centering a mask, sorting the pixels increasingly, and assigning the median value to the kernel pixel. Then, the *bitwise AND* binary operator was applied between the generated image and the original one. The process is shown in bottom part of Fig. 5.3.

Finally, a resize operation was performed for all the images, starting from a resolution of 600×450 pixels to 300×225 pixels. This is performed to reduce the calculation complexity in the following steps.

To partially solve the problem of unbalanced data, a Data Augmentation process was carried out on melanoma images. In fact, through concatenated operations composed of horizontal and vertical flips, rotations of 180, 90 and -90 degrees, each

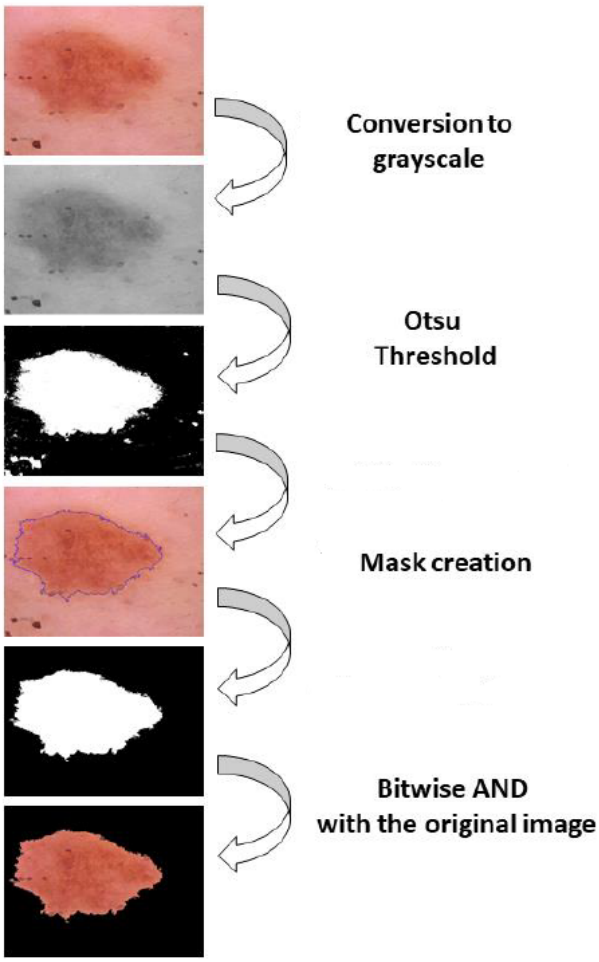


Figure 5.3: Process for the lesion segmentation and extraction.

original image of melanoma has been used to generate seven distinct images. Thus, we have obtained a total of new 7.791 images that, added to the initial 1.113, formed 8.904 images of skin lesions related to melanoma.

There are different techniques to handle unbalanced data. In this work, we combined two techniques: *weight balancing* through a heuristic approach inspired by King *et al.* [124] and *real-time data augmentation in training process* where the images are randomly augmented during execution, therefore at each iteration, the training algorithm will learn from the augmented images.

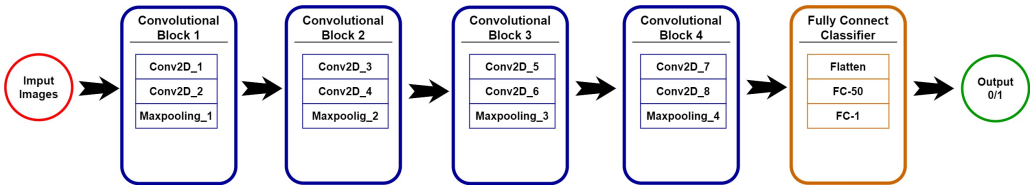


Figure 5.4: The Convolutional Neural Network architecture.

Deep Neural Network Analysis

We have designed, implemented and trained three different deep learning approaches for analysis [82], i.e., a 2D Convolutional Neural Network (2D CNN), a Residual Neural Network (ResNet) and a Self-Organizing Map Neural Network (SOM).we have designed, implemented and trained three different deep learning approaches, i.e., a 2D Convolutional Neural Network (2D CNN), a Residual Neural Network (ResNet) and a Self-Organizing Map Neural Network (SOM).

Convolutional Neural Networks (CNNs) is a family of neural networks widely used in the field of computer vision and, more generally, with data having spatial relationships. It is a system based on a mathematical model, which in addition to identifying features in the images, is also able to assign them a classification, thus giving semantics to the image itself. CNN is made up of modules (levels in which neurons are organized) that can be connected in different ways: the most common is to connect the modules in cascade. Each module takes an input in the form of a volume of pixels organized according to width, height, and depth, and outputs another volume of pixels, or a vector of scores distributed according to one of three dimensions. The input volume to CNN is sized according to the width, height, and depth of the input image, while the volumes produced by the modules have dimensions dependent on those of the input to be processed and other characteristic hyperparameters that vary from one module to another [87]. CNN uses a variable number of learnable filters. The filters in the first levels can understand simple patterns while the following ones succeed in understanding more complex patterns [184]. Finally, the last levels, composed of fully connected neurons, are capable to make predictions.

We used the 2D CNN [145] in order to perform a two-dimensional spatial convolution over images.

The 2D CNN was trained through 100 epochs with a batch size of 50 and using the Adam optimizer. We made up the neural network with 4 convolutional blocks (composed of Convolutional 2D and Maxpooling layers) and one last fully connected (FC) block for the classification. The training set, test set and validation set have been set by considering the percentages 80%, 20% of the initial dataset, and 20% of the train-

ing set, respectively. The convolutional levels use a kernel size of 3 and a stride of 1 and a regularization of the L2 kernel for convolutional and dense layers. We used the Rectified Linear Units (ReLUs) as the activation function for each convolutional level and the Sigmoid activation function in the last FC level to have a binary classification of the problem. We also added two levels of Dropout; one after all the Convolutional 2D and Maxpooling layers equal to 0.25, and the other after the first FC level also equal to 0.25. The architecture of the network is shown in Fig. 5.13.

A Residual Neural Network (ResNet) is a type of artificial neural network that is inspired by pyramidal cells in the cerebral cortex. The central idea of ResNet is that, in the construction of a neural network with a high number of levels, the representation of the input data should remain as unchanged as possible by going deep into the network, so as to preserve information. In deep neural networks, the accuracy of the model increases as the number of layers increases, but there is a limit of the levels to be added in order to have an improvement in the precision and accuracy of the model, since if you continue to increase the number of layers accuracy will reach a point where it will saturate. But deciding the number of layers to add to a network is not so simple, for this reason we used the ResNet, the latter uses a new and innovative type of blocks (called residual blocks) and the concept of residual learning, that is through concatenation of different blocks. ResNet learns to predict a certain output not by learning a direct transformation from the input data to the output, but by learning a certain term $F(x)$ to be added to the input data to arrive to the output minimizing the error. Another component widely used in ResNet is represented by the batch normalization levels [111], used after each convolution and activation. Batch normalization is an operation that allows you to normalize the data present in mini-batches, and thanks to this it reduces the limitations on the value of the learning rate that typically exists in the training of deep neural networks. It also makes the initialization phase of the weights less complex. All this leads to a significant reduction in the time required for the network training process.

ResNet uses skip connections or shortcuts to jump on some layers. The concept on which the residual block is based is to subject an input x to the sequence of convolution-ReLU-convolution operations, obtaining a certain $F(x)$, increased by the same x value. Thus, the output is $H(x) = F(x) + x$, whilst a traditional feed forward CNN practically computes $H(x) = F(x)$ [102].

A motivation for skipping levels is to avoid the problem of escape gradients by reusing activations from previous level until the adjacent level learns its weights. Jumping effectively simplifies the net, by using fewer levels in the initial training stages. This accelerates the learning by reducing the impact of escape gradients, as there are fewer layers to propagate. The blocks of the full network is shown in Fig. 5.5

ResNet has been trained for 100 epochs, with a batch size equal to 32, we used Adam optimizer with learning rate equal to 0.001 and clipnorm equal to 1, and binary cross-entropy as loss function. To reduce overfitting and improve generaliza-

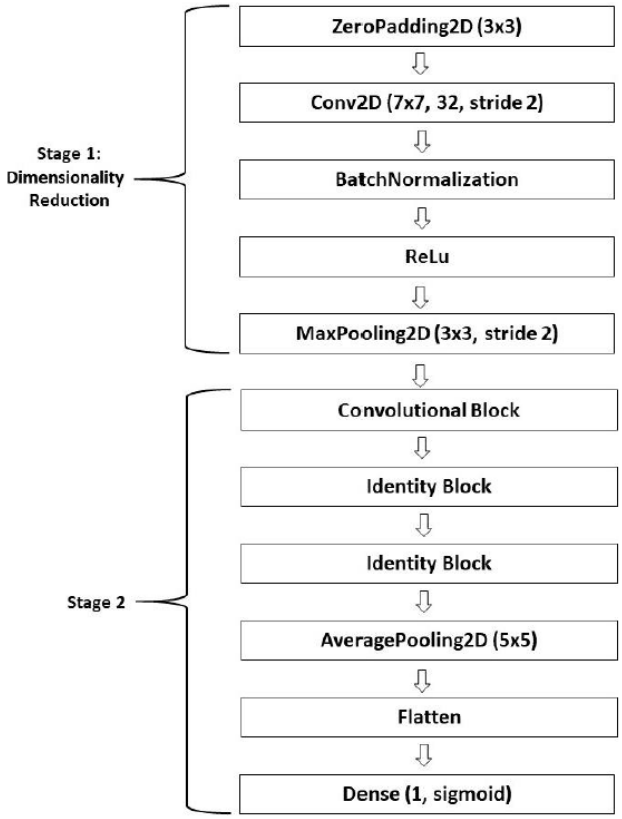


Figure 5.5: The full implementation of the Residual Neural Network.

tion, we used the L2 regularization (Ridge Regression) with lambda parameter equal to 0.0005. Ridge regression adds "squared magnitude" of coefficient as penalty term to the loss function. We used the same percentages of training, test and validation sets adopted for the 2D CNN.

The Self-Organizing Map Neural Network (SOM) is inspired by the cerebral cortex where nearby neurons are activated by similar stimuli, taking into account their connections and the influence that a neuron can have on its neighbors. In general, each neuron is assigned a weight and the ones with similar weights move closer, while those with very different weights move away [2]. When a pattern x is provided to the network, each neuron i calculates the distance from its vector of weights w_i to x. Similarly to the behavior of biological networks, the neuron that with the least

distance from the input pattern has the greatest response by inhibiting the adjacent neurons and it is charged to update the weights [127].

The SOM we used, namely $SUSI$, was a neural network that applies a combination of unsupervised and supervised learning, in order to perform classification on the images taken as input [190]. The main phases of SUSI are shown in Fig. 5.6.

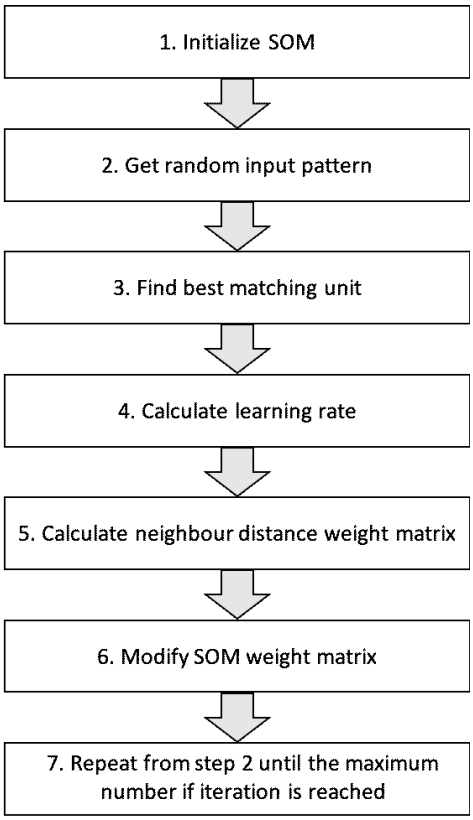


Figure 5.6: The main phases of SUSI.

We initialize the SUSI randomly. During the search for the best matching unit, the current input pattern is compared to all n-dimensional weight vectors on the SOM grid. The SOM node that is the closest one to the input node according to the chosen distance metric is the BMU. The learning rate of the SOM is a function that decreases when the number of iterations increases. This implies a faster convergence and prevents oscillations. Similar to the learning rate, the neighborhood function is monotonically decreasing. The neighborhood distance weight is a func-

tion of the number of iterations and the distance $d(BMU, i)$ between the BMU and every other node i on the SOM grid. The distance d is defined as the Euclidean distance on the 2D map grid. After reaching the maximum number of iterations, the unsupervised SOM is fully trained [191].

The number of neurons used to configure the grid was 20×20. The initial learning rate used was 0.5 while the final rate was 0.01. The training was carried out using a number of supervised/unsupervised iterations both equal to 2000. Subsequently the model was built on the training set, tested on the test set and validated on the validation set. We used the same percentages of training, test and validation sets adopted for the 2D CNN.

Results

To compare the impact of pre-processing phase on the classification, we used two versions of the dataset: the original one and a pre-processed version. To test results we calculated the accuracy, sensitivity and specificity for each adopted neural network model:

- *Sensitivity:* it represents the ability of the test to correctly identify the diseased state, and it is the proportion of true positives that are correctly identified by a diagnostic test:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.1}$$

  where TP represents true positives and FN represents false negatives.

- *Specificity:* it is the ability of the test to correctly diagnose the benign cases, it is the proportion of true negatives correctly identified by a diagnostic test:

$$Specificity = \frac{TN}{TN + FP} \tag{5.2}$$

  where TN represents true negatives and FP represents false positives.

- *Accuracy:* it indicates the proximity of the value found to the real one, and it is computed as the number of correct predictions (true detected cases) divided by a total number of predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.3}$$

The results obtained without the pre-processing phase are shown in Table 5.1, whilst the ones obtained by considering the pre-processing are shown in Table 5.2. In particular, there is an improvement in terms of accuracy in the range $(2, 2-2, 5)$,

Table 5.1: The neural networks results **without** pre-preprocessing.

	Accuracy	**Sensitivity**	**Specificity**
2D CNN	71,9%	69,0%	**92,8%**
ResNet	**79%**	**80%**	78%
SOM	66,8%	61,7%	63,5%

Table 5.2: The neural networks results **with** the pre-processing.

	Accuracy	**Sensitivity**	**Specificity**
2D CNN	74,1%	**89,4%**	72,1%
ResNet	**81,5%**	85%	**79%**
SOM	69%	64%	68%
Dermatologists	80%	82% (68%-98%)	59% (34%â72%)

i.e., 2,2% for 2D CNN, 2,5% for ResNet and 2,2% for SUSI neural network. Moreover, for the 2D CNN without pre-processing, we obtained a high specificity value (i.e., 92,8%) with respect to the sensitivity one (i.e., 69%). This highlights a worst behaviour of the network in determining the melanoma cases. On the contrary, for the same neural network, we obtained opposite results with the pre-processing (72,1% for specificity with respect to 89,4% for sensitivity) by highlighting a better behaviour in classifying the melanoma disease.

We have obtained a better result with ResNet (i.e., 81,5% for accuracy) because in 2D CNN (with an accuracy of 74,1%) there is generally a problem of cancellation of the gradient, whose descent, given by the minimization of the error function, is reduced exponentially through the backpropagation of the previous layers, what with the ResNet does not occur due to residual learning. While for SOM (69% of accuracy), the main limitations is the static nature of the size and topological structure of the output grid. The SOM starts from a preset graph whose vertices are placed on the data space not ensuring the construction of a Topology Preserving Map. This aspect has a particularly negative impact in the case of incremental problems, where it is also desirable that the network structure and its size may evolve with the progress of the acquired data. This issue is confirmed by the U-Matrix calculated on the best of the tests carried out with the SOM. The U-Matrix provides a clear view of the results through a visual representation of the distances between neurons in the input data dimension space. Unfortunately, the U-Matrix depicted in Fig. 5.7 shows the presence of darker colors representing high distances among the neurons. This means bad distribution of data on the neurons.

All possible combinations of sensitivity and specificity make up a ROC space. A

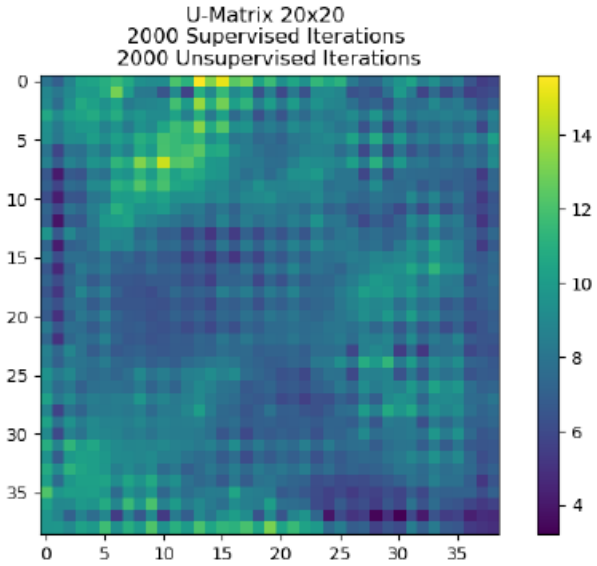


Figure 5.7: U-matrix: visual representation of the distances between neurons in the SOM network.

single point in the ROC space shows the compromise between sensitivity and specificity, i.e., the increase in sensitivity is accompanied by a decrease in specificity. In order to evaluate whether the diagnostic classification is good or not, we have analyzed the ROC curve of the best model, i.e., the ResNet as shown Fig. 5.8. In particular, the area under the curve (AUC) value greater than 0,80 is indicative of good discrimination, whereas a value less than 0,70 represents a insufficient discrimination [105]. The ResNet obtained an AUC equal to 0,875.

Finally, we compared our results, in terms of sensitivity and specificity, with a study carried out chosen from 12,000 dermoscopic images in the ISIC archive. The dermatologists achieved a mean sensitivity of 82% (range: 68%-98%) with a mean specificity of 59% (range: 34%-72%). For accuracy we have compared our results with [161] where dermatologists have achieved an accuracy of 80% in diagnosing malignant lesions. From these results, our analysis has obtained a better accuracy than dermatologists with the ResNet, a greater sensitivity with the 2D CNN and the ResNet and a greater specificity with all the neural networks used.

The final accuracy obtained of our proposal is quite similar to the ones listed in Table 3.1, whilst it slightly outperforms the ones that have been applied on the ISIC

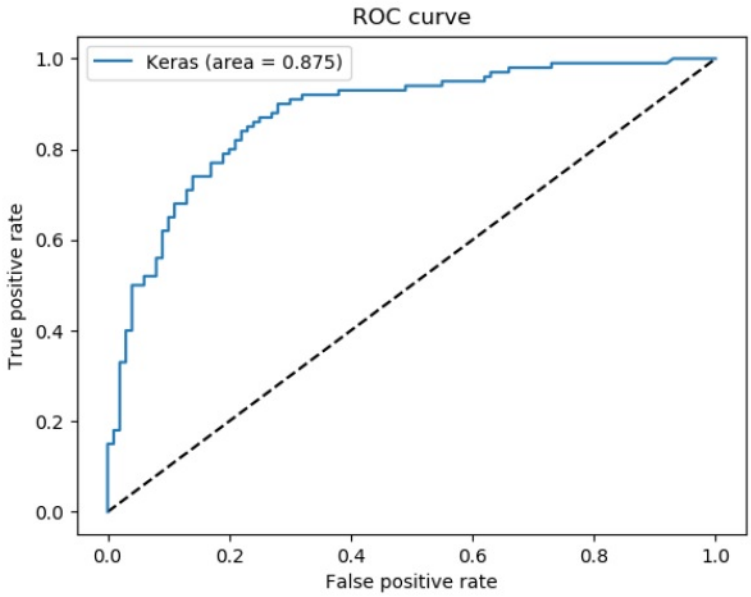


Figure 5.8: The ROC curve of the ResNet model.

archive.

5.1.4 An Augmented Reality Mobile Application for Skin Lesion Data Visualization

In this section I describe the proposal of an Augmented Reality smartphone application for supporting the dermatologist in the real-time analysis of a skin lesion. The app augments the camera view with information related to the lesion features generally measured by the dermatologist for formulating the diagnosis. The lesion is also classified by a deep learning approach for identifying melanoma. The real-time process adopted for generating the augmented content is described. The real-time performances are also evaluated and a user study is also conducted. Results revealed that the real-time process may be entirely executed on the Smartphone and that the support provided is well judged by the target users. The approach supports the clinician in diagnosis of melanoma by using AR visualization technique on a mobile device and deep learning classification. To the best of our knowledge similar approaches are still lacking in the literature. We also consider other rules in addition

to ABCD for decision support, including a new technique (i.e., photometric stereo) never adopted for melanoma feature extraction.

Rules for Melanoma diagnosis

The most adopted rules for the early Melamoma diagnosis are yhe ABCDE rules, based on the extraction of five characteristic features of this tumor, as shown in Fig. 5.1[2].

(A) Asymmetry: A benign nevus is generally circular or in any case round, melanoma is more irregular and larger; for this reason, the asymmetry of a lesion is of fundamental importance in diagnosis. The lesion is generally contained in a larger spot. Dermatologists generally evaluate asymmetry by comparing the two halves of the lesion according to its main axis. For computational detection of asymmetry, we adopted an algorithm capable of indexing the symmetry of the lesion [212]. This index is assessed by defining a percentage of lesion asymmetry according to the main axis, i.e., by evaluating different parameters, including the eccentricity of the position, calculated as follows: $A = \frac{\text{Dist}}{\sqrt{\text{Area}}}$, where $Dist$ is the Euclidean distance between the centroid of the largest spot and the centroid of the lesion and $Area$ is the area of the lesion.

(B) Border or lesion segmentation: The process of separating the lesion from the surrounding skin to isolate the region of interest is based on edge detection or image segmentation techniques. The approach is the skin segmentation adopted during the early phase of the CNN.

(C) Color detection: It is very important as the light brown, dark brown, black, red areas signal of vascularized areas and could be indicative of malignant lesions [222], such as: *i) White/bluish veil* (WB) often present in melanoma, rarely in Spitz/Reed nevi. *ii) Bluish gray areas* (BG) can be associated with melanoma in regression and can appear as areas scattered with peppering and granules or be well defined with irregular edges. *iii) Blackheads and brown blood cells* (BB) correspond to accumulations in the papillary dermis or to the junction of pigmented melanocytes or melanophages; they are rounded spherical bodies of variable color. If arranged in the suburbs (mainly in adults) and irregular in shape and size, they can be a sign of suspected injury. Red and white blood cells are present in melanoma and are due to increased vascularization. *iv) Diffuse pigmentation* (DP) is related to irregular localized areas can be significant for melanoma. *v) Depigmentation* (DE) can be a sign of a regressive phase of melanoma and correspond to areas of fibroplasia, telangiectasias, and loss of melanin. *vi) Miliary cysts and similcomedonic outlets* (MC) are roundish whitish or yellowish areas and are often present in seborrheic keratoses, sometimes they are also present in melanoma and melanocytic nevi.

[2] http://clovisdermatology.com/melanoma

Common situations involve the use of RGB images that can be managed through the individual red, green and blue color channels. However, the most common in melanoma, can also be managed through the use of Hue Saturation Brightness (HSB) also called HSV from Hue Saturation Value, an additive color composition method [227], where H, S and V are defined as follows (in terms of RGB color):

$$S = 1 - \frac{3}{(R+G+B)} \cdot [\min(R,G,B)],$$

$$V = \arccos\left\{ \frac{R - \frac{1}{2}(G+B)}{\sqrt{[(R-G)^2 + (R-B)(G-B)]}} \right\}$$

$$H = \begin{cases} W & if\ G > B \\ (2\pi - W) & if\ G < B \\ 0 & if\ G = B \end{cases}$$

(D) Diameter: The size of a lesion is very important for the detection of a malignant lesion. Generally, a suspect lesion has a diameter $> 6mm$ [103], but recent studies have shown that, with increasing melanoma diameter, the Breslow depth also increased and that about 30% of lesions less than 6 mm were invasive [220]. For detecting the diameter we considered the center of gravity of the lesion, a method also used to calculate the asymmetry index for the differences between the areas [36].

(E) Evolution: The evolution of a lesion consists in the modification of size, shape and color in a rather short period of time, approximately 6-8 months. In the present version of our system we have not considered this parameter yet.

In addition to the ABCDE rules, we introduce the visual analysis of a peculiar characteristic of melanoma that should be considered to obtain a correct classification, the *Palpable elevation of the lesion.* This characteristic is evaluated by the clinician, but (for the best of our knowledge) is not considered yet in software tools. The appearance of a papule or lump in the context of a pigmented lesion can often denote the presence of a malignant lesion. A *Photometric stereo* algorithm has been adopted to measure the degree of elevation of the lesion. Photometric stereo is a method to estimate depth and surface orientation from images of the same view taken from different directions (used in many 3D reconstruction applications) and is based on normal constructions based on the direction of light. Generally, three directions are enough to get the normals, but a larger number is needed to minimize the process noises. First, the direction of the light is computed. Lambertain surfaces are used for the generation of normal maps, where the intensity at any point on the surface can be given as $I = \frac{N \cdot L}{\pi}$,where L is the direction of the reflected light, N is the normal at the surface point and and is computed by considering at least three light sources that are not in the same plane based on the direction of the x, y and z axis [16]. Finally, the 3D reconstruction is mapped into a 2D image.

We also analyze the *Pigmentary reticulum and pigmentary pseudoreticle* features. The former is the most important parameter, the lines of the network correspond to the elongated epidermal crests and the spaces of the network to the dermal papillae; in benign lesions it appears regular and nuanced in the periphery, while in suspicious lesions it appears irregular with coarse meshes, with pigmentation of varying intensity, unshaded in the periphery and asymmetrical [116]. The Pigmentary pseudoreticle it is an interruption of pigmentation by hypopigmented patches determined by hair follicles and glandular outlets. In lentigo maligna due to the atypical melanocytes increased in number, it appears irregular and coarse [14]. In our proposal, we use the *Fractal dimension* (F_d), which measures the repetition of each sub-structure at a specific scale applied to the image. F_d provides the N_r distinct (non-overlapping) copies of each sub-structure scaled by a ratio r (where $0 \leq r < 1$ when the image is scaled down) [38]. The fractal dimension F_d is given by the relation:

$$1 = N_r r^{F_d} \implies F_d = \frac{log\,(N_r)}{log\,(\frac{1}{r})}$$

The proposed Augmented Reality app

The mobile system we propose aims at supporting the clinician in the analysis of melanoma lesions by visualizing in augmented reality modality information provided by the feature extraction phase described in the previous section, as shown in Fig. 5.9. It is based on a the following phases:

- *Distance detection.* The distance between the mobile device and the patient skin is detected. The optimal distance is set to 10 centimeters.
- *Skin lesion centering and selection.* By analyzing the screen image, the system is able to detect if a lesion skin is in the center, resulting inactive until the dermatologist focus a lesion in the middle of the screen. In the case of separated multiple skin lesions present on the center, it allows the dermatologist to focus by selecting one of them (the system suggests the closest to the center), as shown in Fig. 5.10(a).
- *Optimal light detection.* The optimal light is detected in order to better represent and visualize the lesion skin. It is performed by analyzing the skin lesion and detecting the presence of white reflections (due to the existence of not removed gel applied to clean the nevi). Moreover, in the case the environmental light is not enough or the image is blurred, the system turns on incrementally the flashlight of the smartphone until the intensity of the light is optimal and the image of the nevi results unblurred.
- *CNN classifier and feature parameters.* The system computes all the features described in the previous section on the selected lesion. During this phase, the image of the lesion is preprocessed and provided as input to the CNN classifier.

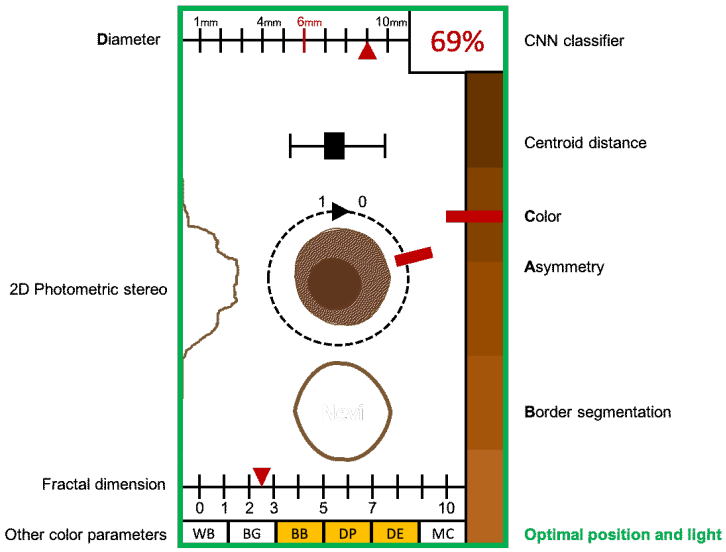


Figure 5.9: The main visualization layout.

- *Building the AR visualization.* The system integrates all the collected information, builds and shows the augmented visualization on the mobile screen (see Fig. 5.10(b)).

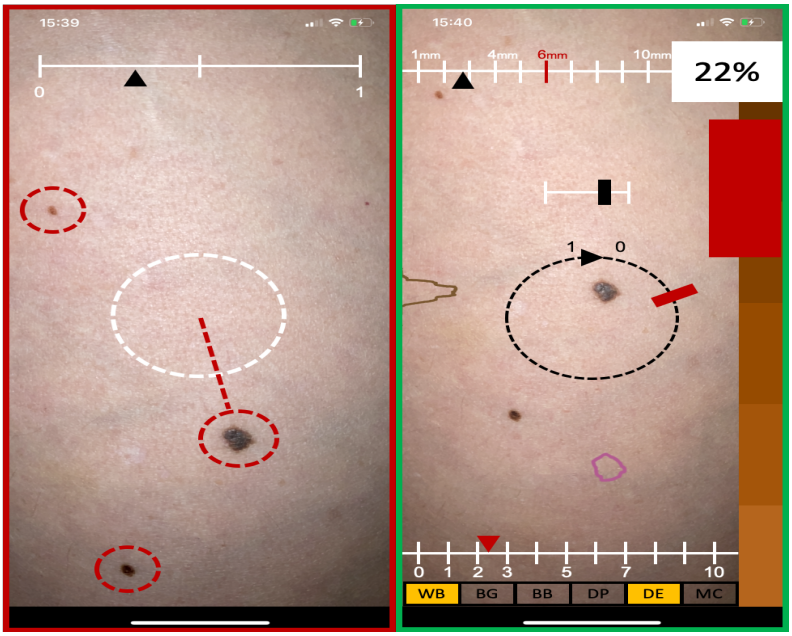
The AR mobile application has been developed and tested on an iPhone 11 Pro Max, with software version 13.3.1 and ARKit framework 3.5 (see Fig. 5.10).

In a preliminary evaluation, the mobile application has been used by 7 dermatologists (2 of them were against the use of technologies and image processing). However, 6 of 7 provided positive and very positive feedback about the usefulness of the provided visualization approach.

- To analyze a nevus the dermatologist has to frame the lesion on the patient skin by using the device camera.

- To detect the distance between the patient skin and the smartphone, which should be set to 10cm.

- To analyze the image signal if the skin lesion is in the center of the device display. When there exist several skin lesions in the center of the screen, the app highlights them, as shown in Fig. 5.10(a) and enables the clinician to select the one to be examined among them. The more central one is suggested.



(a) (b)

Figure 5.10: Select and center a nevi (a), the AR visualization (b).

- Once a nevus has been individuated the system has to keep on tracking it until it is not visible anymore.
- Classify an image and compute all the features to support the diagnosis in real-time.
- To integrate and pose the measures in AR on the mobile screen.

In this research [81], we focus on the collection of the skin lesion information and on the tracking of the dermatologist camera which has to overlap the support information to the camera view in a time little enough to avoid an inappropriate waiting. When the user moves the smartphone all the information has to be recalculated and updated.

Video-based AR captures the camera video channel which is adopted as the background of the image. This image is then added it to the video of the camera. To this aim the system has to map the position of the AR contents in real-time.

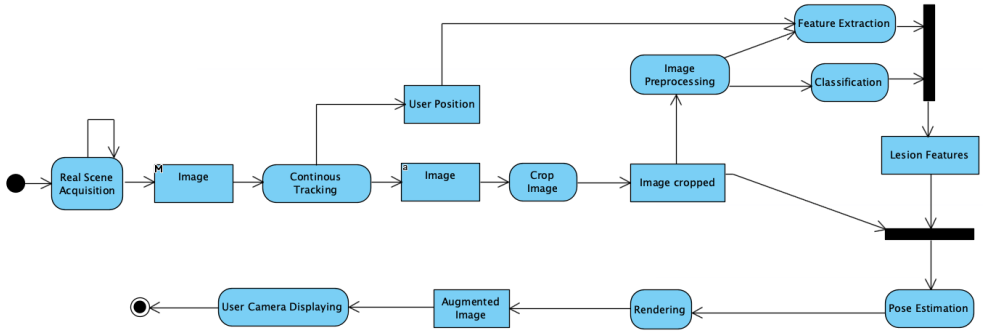


Figure 5.11: The real-time process.

The real-time process

In the last years, the computational power of smartphones and tablets is largely increasing, giving them the power which was available on PC not long ago. Thus, the former may run without problems the most part of mobile applications, while some functionalities may be challenging, such as AI algorithms [110]. Thus, different client-server architectures may be adopted, depending on how we divide the computation between the client and the server. Initially, we decided to use top-level smartphones and leave all the computation on them. Then the architecture may be modified depending on the results of the real-time performance evaluation.

The real-time skin analysis process is based on the following phases, as shown in the activity diagram with object-flow depicted in Fig. 5.19:

The following fundamental steps have to be performed to provide the appropriate support:

1. Real-time Acquisition. A frame of the camera video is acquired.
2. Continuous Tracking. The system has to track the device's position with respect to the patient skin. This is useful to determine the distance and the device orientation. The nevus position is also determined.
3. Image Crop. The app automatically crops the nevus.
4. Image Pre-processing. The nevus image is pre-processed for maintaining only the relevant aspects of a lesion.
5. Feature extraction. ABCD rules, 2D Photometric Stereo, and Fractal Dimension are computed.
6. Classification. CNN performs a classification of the nevus.

7. Pose estimation. It computes the position in which the contents have to be added on the camera view with respect to the nevus position.

8. Rendering. The evaluated features are combined with the original image.

9. Displaying. The augmented image is shown in the dermatologist camera view.

Continuous Tracking. To display information in real-time on the user camera there is the need of tracking the patient nevus while the dermatologist moves the device and determining the device's distance from the user skin. One of the main concerns to get consistent images are the lighting and the patient positioning blur [10]. The application suggests putting the point of interest in the middle of the camera view, as shown in Fig. 5.10(b). This to avoid confusion as in the case of Fig. 5.10(a), where the image contains two nevi.

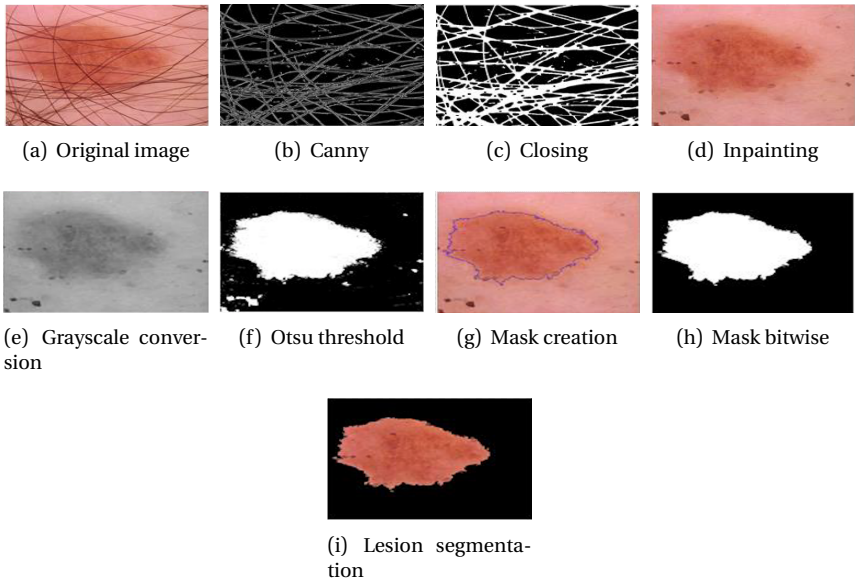
The distance between the device and the patient skin is measured by following the method proposed in [134]. In particular, this method uses the object disparity, dragging distance, and device orientation to determine the distance between the patient skin and the device. To measure the distance, users have to hold the mobile device in upright position and drag it along the vertical direction (perpendicular to the patient skin). The algorithm is based on the relation between object disparity from two images and a difference in camera positions. Acceleration signals during the dragging period are analyzed to compute the dragging distance. See [134] for further details. The application produces a red circle around the nevus when the distance is not appropriate. The circle is black when the distance is about 10cm. The distance is successively used to compute the size of the nevus.

To identify the nevus we adopted the motion tracking approach proposed in [244], where the nevus feature points taken in the first frame are tracked in the successive frames.

Image Feature Detection. Before using an image for extracting the relevant features of a lesion or to provide it as input to a classifier there is the need to pre-processing it. This activity is performed in the following three phases:

- *Hair removal.* The presence of hairs that occlude a skin lesion may interfere with its diagnosis. We used the *Canny edge detection* algorithm for the hair removal and *Otsu's threshold* has been used as a mask for hair removal. Finally, we performed image inpainting, a technique that performs a sort of interpolation for digital image processing to reconstruct parts of damaged digital images [51]. The hair removal process is shown in Fig. 5.12(a)-(d);

- *Lesion segmentation.* Segmentation aims at selecting specific objects or regions in an image based on a choice of properties, such as brightness, color, and texture. The image is converted from RGB to grayscale and the lesion is separated from its background (i.e., the skin). The Otsu's Thresholding method is used to automatically perform clustering-based image thresholding.



(a) Original image (b) Canny (c) Closing (d) Inpainting

(e) Grayscale conversion (f) Otsu threshold (g) Mask creation (h) Mask bitwise

(i) Lesion segmentation

Figure 5.12: The skin lesion image processing.

- *Clinical feature segmentation.* In this case, median filtering was used which preserves the edges by removing the over-segmentation and also the blurring effect. Then, the *bitwise AND* binary operator was applied between the generated image and the original one. The process is shown in bottom part of Fig. 5.12.

Nevus Feature Extraction. In the following, we describe the nevus features to be displayed in AR. We visualize information related to the ABCD features, the Palpable elevation of the lesion and pigmentary reticulum and pigmentary pseudoreticle features.

Melanoma Classification with Deep Learning. To classify melanoma we adopted a Convolutional Neural Network (CNN). To train the CNN model we adopted a dataset composed of clinical images taken by top-level smartphones. Several kinds of lesions have been collected, including Actinic keratoses, Bowen's disease, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi and vascular lesions.

We decided to consider a binary classification problem (i.e., melanoma or not melanoma). The CNN was trained through 100 epochs with a batch size of 64 and using the Adam optimizer. We made up the neural network with 4 convolutional blocks and one last block for the classification. The training set, test set, and validation set have been formed by considering the percentages 80%, 20% of the initial dataset, and 15% of the training set, respectively. The convolutional levels use a kernel size of 5 and a stride of 1 and a regularization of the L2 kernel. We used the Recti-

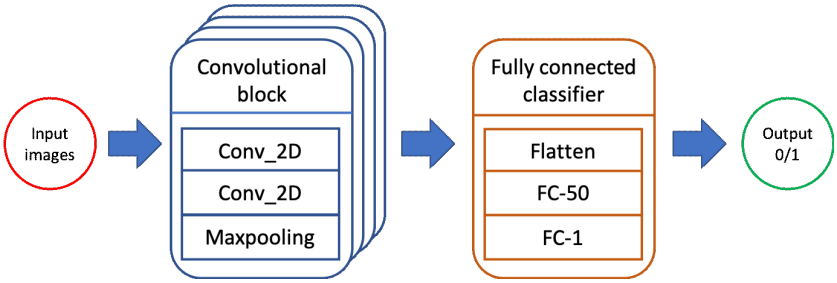


Figure 5.13: The Convolutional Neural Network architecture.

fied Linear Units (ReLUs) as the activation function for each convolutional level and the Sigmoid activation function in the last fully connected level (FC) to have a binary classification of the problem. The architecture of the network is shown in Fig. 5.13.

Table 5.3: The CNN classification results.

Accuracy	Sensitivity	Specificity
78.8%	**91.3%**	73.0%

Classification Results. To analyze the classification results of the adopted CNN model we computed accuracy, sensitivity, and specificity (see Table 5.3). The CNN classifier obtained an accuracy average result of 78.8%.
Pose Estimation. In this phase, we determine the circle position in AR given the nevus position in the frame, as shown in Fig. 5.14. This is simply performed by considering the diameter dimension and the center of gravity of the lesion.
Rendering. In this phase, the app combines the augmented content and the original image on the base of the pose estimation results. In particular, the circle around the nevus has to be positioned according to the coordinate computed in the camera pose phase together with the nevus features and the classification result, as shown in Fig. 5.10(b).

Usability study

Several participatory design groups involving clinicians and researchers have been conducted before designing the application prototype analyzed in this usability evaluation. First, a mock-ups prototype has been created (a screen is shown in Fig. 5.9); the mocks-ups were evaluated and then the running prototype has been implemented.

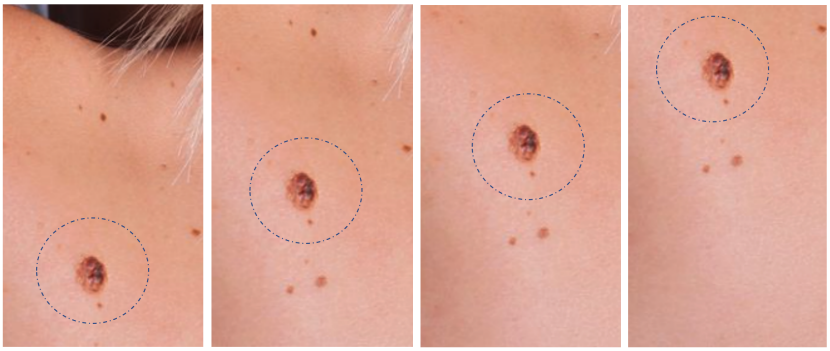


Figure 5.14: Nevus tracking.

The *goal* of this study is to *evaluate* usability and the user satisfaction of the proposed app in a clinical setting from the *point of view* of professional dermatologists.

Six professional dermatologists voluntarily participated in the study. They were 35-60 years old and were frequent smartphone users and were accustomed to using them to take pictures of patient nevi (3 of them were male). Each one of them individually experimented in his office, adopting think-aloud protocol and under the supervision of one of the authors (the same in each evaluation). Before starting the experiment the supervisor performed a 5-minute presentation showing the app's goal and operation. Participants (dermatologists and patients) provided written informed consent and were informed that their data and multimedia content were managed in an anonymous form. We asked the participants to follow the thinking aloud protocol as they were registered.

Tasks. Each dermatologist examined the skin lesions of three new patients by performing the scenario composed of the tasks reported in Table 5.4 for each suspect skin lesion. The scenario was derived by the requirement analysis document of the application and concerned the most relevant use cases.

The videos of the patient skin were recorded and used in the real-time evaluation of the system described in the following.

Variables and Materials. We considered the following variables: the scenario completion success rates, the time (expressed in seconds) needed to accomplish the task, the number of errors while performing a task, such as navigation errors; presentation errors (e.g., failure to find and properly act upon the interface element), selection errors due to labeling ambiguities; control usage issue (wrong use of entry field). These errors were noted by the supervisor.

The dermatologist perceptions were collected at the end of the experiment through the Post-Experiment questionnaire in Table 5.5. In particular, the P0 question re-

Table 5.4: The tasks composing the scenario of use.

Task	Description
T1	Start the mobile application
T2	Logging in
T3	Enter Patient Data
T4	Perform the skin evaluation
T5	Examine the lesion parameters
T6	Enter the lesion evaluation
T7	End Evaluation

ferred to the clearness of the task to perform. The participant perception of the system Usability has been collected by using the standard Italian version of the System Usability Scale (SUS) questionnaire [29], is a Likert Scale which consists of 10 questions (P1-P10 questions in Table 5.5). Each question is ranked from 1 (disagree vehemently) to 5 (strongly agree). We also add three questions with the same Likert scale: P11 and P12 for explicitly collect perceptions on the loading time and the visual metaphor of the augmented content during the nevus analysis, respectively; P13 (overall) resumes the participant opinion on the support offered by the tool and P14 collects open comments.

Table 5.5: Post-Experiment questionnaire.

ID	Question
P0	The tasks to perform were clear.
P1	I think that I would like to use this system frequently.
P2	I found the app unnecessarily complex.
P3	I thought the app was easy to use.
P4	I think that I would need the support of a technical person to be able to use this app.
P5	I found the various functions in this system were well integrated.
P6	I thought there was too much inconsistency in this app.
P7	I would imagine that most people would learn to use this system very quickly.
P8	I found the app very cumbersome to use.
P9	I felt very confident using the system.
P10	I needed to learn a lot of things before I could get going with this system.
P11	Is the loading time of the augmented skin lesion information during the lesion analysis satisfactory.
P12	The metaphors for depicting the skin lesion features during the lesion analysis are easy to understand.
P13	Overall, I'm satisfied of the support offered by the tool in the skin lesion examination.
P14	Open comments

The challenge of an AR application on a mobile device is speed [228]. In particular, the main requirements of our application are:

- performing the tracking process in a short time;

- computing the lesion features and classification in a short time;
- providing the augmented information back to the dermatologist in a short time.

The proposed approach was tested by creating video test sequences at a 2160p video at 60fps and 30fps, the video format of Samsung S10. We extracted from the videos produced during the usability evaluation 12 sequences (two for each dermatologist): 6 with melanoma, 6 with not pathological nevi, having a length ranging from 710-1320 frames. Following the direction traced in [228], we compared the proposed approach performance on a mobile device and on a desktop in such a way to be able to decide which activities shown in Fig. 5.19 may all be performed on the client or on the server. To this aim, there is the need of providing the video stream to both systems. Thus, we developed a frame server loading a video previously taken by the dermatologist mobile device from the file system of both the devices in substitution of the mobile camera. Because performances scale linearly with the CPU clock rate on mobile phones they do not dependent on the operating system [229], we performed our benchmark by using a single device. The Samsung S10 mobile device was equipped with a Qualcomm Snapdragon 855 processor, 1785 MHz, which allows the running of trained neural networks on the device without a need for connection to the cloud [183]. The MacBookPro (2019) had a processor with 2.6GHz 6-core Intel Core i7, with 12MB shared L3 cache, 512GB SSD, 16GB of 2400MHz DDR4 onboard memory.

The CNN model was trained on a Desktop PC and then uploaded on the Android smart-phone. In particular, Qualcomm enables us to convert a pre-trained model into the Deep Learning Container (DLC) format. The Qualcomm Neural Processing Engine (NPE) runtime then executes the neural network.

We adopted the TensorFlow[3] framework for the model training because it exposes C/C++ API compatible with the Android platforms.

The times which have a greater impact on the performances are related to the following activities:

- Initialization Time. This is the time intercurring between the pressing of the app icon on the mobile screen and the moment in which the app is ready to accept user input.
- Tracking Time. The tracking time is the time needed to compute the user position and the nevus feature points.
- Pre-processing Time. This is the time taken to perform the pre-processing phase described in Section 5.1.4.

[3] https://www.tensorflow.org

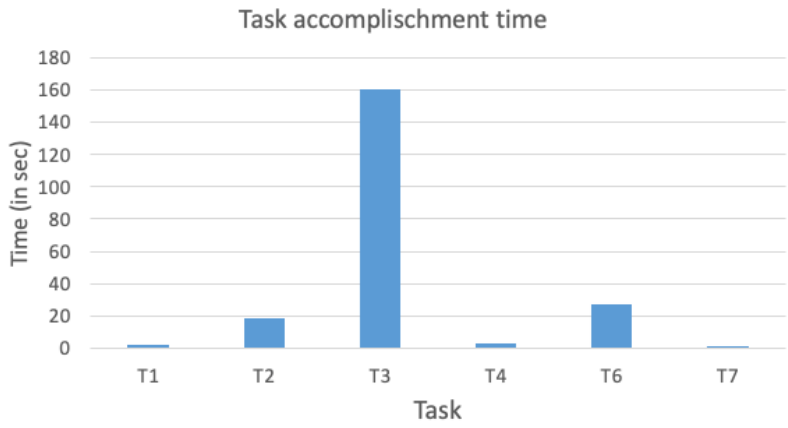
- Classification Time. This is the time the server takes to classify the skin lesion, as discussed in Section 5.1.4.
- Feature Computation Time. The time the server takes to the lesion features described in Section 5.1.4.
- Pose Estimation Time. The time needed to compute the AR content position.
- Rendering Time. The time to superimpose the augmented content to the original frame.

Usability results The scenario completion success rate was 100%. All the scenarios were completed by all the dermatologists with all the patients. The time needed to accomplish each task is reported in Fig. 5.15(a), while the number of errors related to each task was summarized in Fig. 5.15(b). The information related to usability perception and dermatologist satisfaction is depicted in Fig. 5.16.

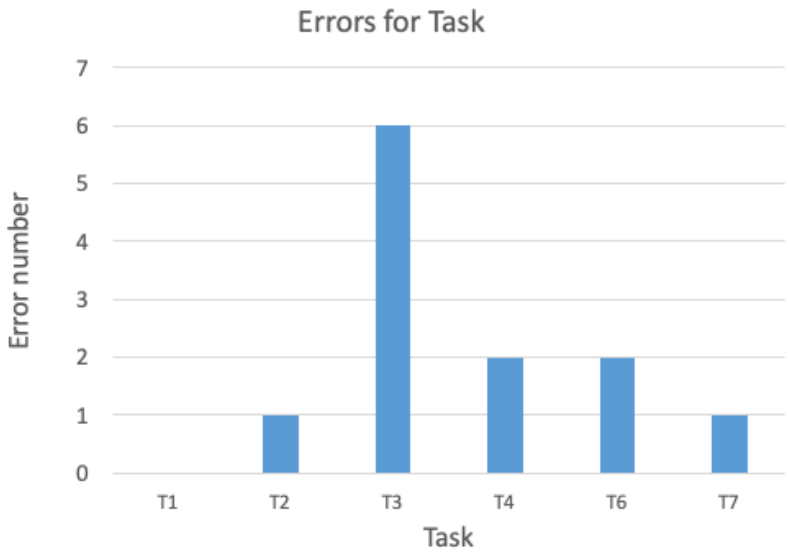
Time. The average time spent to perform the various tasks is summarized in the histogram in Fig. 5.15(a), except for the time related to T5 (examine the lesion parameter), which depends on the lesion complexity and of the dermatologist need of observing it and by the augmented contents. The analysis of the time of the real-time process related to task T5 is reported in the next section. In particular, 2.3 seconds are needed for starting the application. The longest time is taken for filling in the form concerning patient data.

Errors. The number of errors was very reduced. It may be because the users were all smart users. There was some navigation problem in T2 and T6, for accessing the functionalities a wrong button was pressed. The main problem was with task T3, enter the patient data. The main problems were with the data fields, as the supervisor referred.

Usability. The user perception results collected by the Post-Test questionnaire in Table 5.5 are summarized by the histograms in Fig. 5.16, where a histogram is associated with each question. In particular, the tasks to perform were clear for all the dermatologists (**P0**), all of them also think to use the app frequently (**P1**), of which three strongly agreed. The app was not considered excessively complex (**P2**). The app was easy to use for five of them (**P3**) and the other was neutral. All of them agree that there is no need of support for using the system (**P4**) and the system functions were well integrate (**P5**). No inconsistencies were perceived (**P6**), while the app was perceived as easy to learn for 5 participants, one neutral (**P7**). The system was not perceived as cumbersome (**P8**), while all the participants agree that they were confident when using it (**P9**). They also considered unnecessary to learn a lot of things to be able to use the system (**P10**). Both **P11** the loading time and **P12** the adopted AR metaphors are satisfying for 5 participants. The overall satisfaction on the support provided to the skin lesion analysis is very positive for 5 participants, positive for one of them.



(a)



(b)

Figure 5.15: Average time for task (a) and number of errors for task (b).

The same user was neutral in P3, P11, and P12. He explained in the open comments section that *"The bottom ruler meaning is not clear to me. I had some difficulty in stopping the skin analysis."* Another participant lamented the difficulty in remembering the meaning of the same ruler. Thus we decided to activate an information

pop-up when the user touches an element of the interface and to express the symmetry with a numeric value (1, perfectly symmetric, 0 - totally asymmetric).

An appreciation in P13 was addressed to the 3D representation of the lesion, which was considered particularly useful. The same participant suggested inserting a "start analysis button" in such a way the dermatologist may observe the lesion through the device and have feedback from the system only when he needs it. We considered this suggestion very useful and decided to add it to the next prototype version.

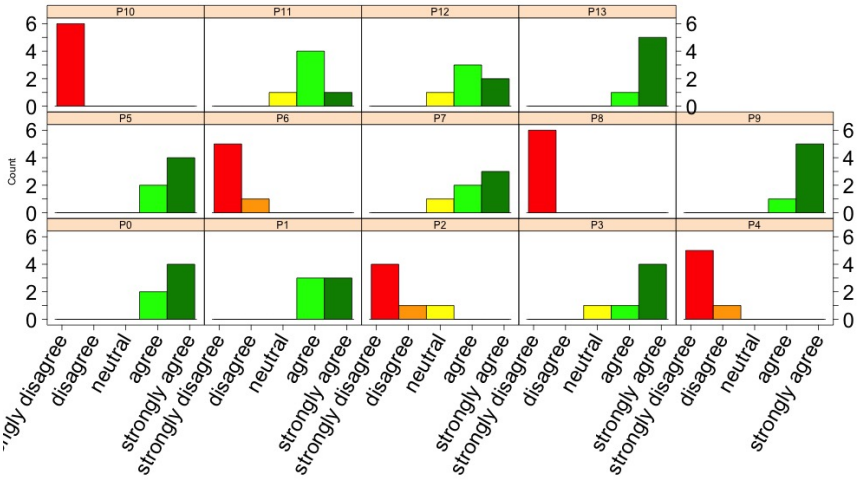


Figure 5.16: The Post-Experiment questionnaire results.

The results in terms of performances of the mobile device concerning the laptop are reported in Table 5.6. We can see that on the mobile device a cycle of the real-time process takes roughly 5 seconds.

5.2 Leukemia classification

In this section I present a deep learning and genetic algorithm based feature selection processes on Leukemia Data for a dataset of patients affected by leukemia.
The patients belong to two distinct classes: Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). Each of them is characterized by a list of identical genes for all the patients. The analyzed data are extracted from dual-channel microarray experiments from the Gene Express Omnibus (GEO) platform, a public database available on the NCBI website containing genomic data, which represent

Table 5.6: Average process performance measures.

Activity	Mobile	Desktop
Tracking time	38 ms	3.8 ms
Image Pre-processing	1.51 s	1.11 s
Feature Extraction	0.9 s	0.35 s
Classification	2.2 s	1.5 s
Pose Estimation	340 ms	9 ms
Rendering	11 ms	1.9 ms

the methylation values for each gene of each sample. The analysis exploits feature selection techniques aimed at reducing the consistent number of variables (genes). To this aim, we use linear models for differential expression for microarray data, and an autoencoder based unsupervised deep learning model to simplify and speed up the classification. Following the reduction in the number of variables, classification models have been implemented with the use of a deep neural network (DNN), obtaining a classification accuracy of approximately 92%. Then, the results have been compared with the ones provided by an approach based on support vector machines (SVM) giving an accuracy of 87,39%. Moreover, another feature selection approach based on genetic algorithms has been experimented obtaining 60,36% (DNN) and 30,63% (SVM) of accuracy. For further verification of the relevance of the selected set of genes, we conducted a gene enrichment analysis based on the functional annotation of the differentially expressed genes. As a result, a differentially expressed pathway between the two pathologies has been detected. In the following the study is deeply detailed.

In the following I introduce the main concept related to Leukemia, the dataset adopted and the proposed analysis processes that are also compared and evaluated.

5.2.1 Leukemia

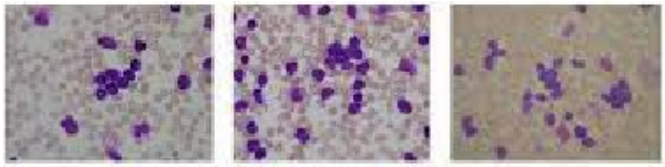
Leukemia is one of the most relevant: in 2020 worldwide it has been the cause of death for 311,594, while the new leukemia cases have been 474,519[4]. For leukemias we mean a heterogeneous group of neoplastic diseases, which foresee any process of proliferative alteration of a progressive and irreversible nature of the blood cells of the bone marrow. Leukemias originate from the malignant transformation of

[4] https://gco.iarc.fr/today/data/factsheets/cancers/36-Leukaemia-fact-sheet.pdf

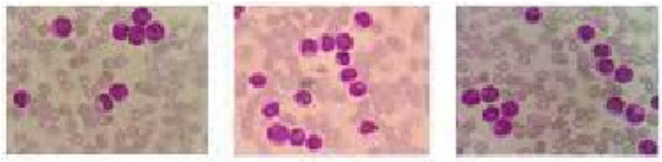
hematopoietic stem progenitor cells, with alteration of the proliferation and differentiation of the same cells (examples are in Fig. 5.17).

In leukemias, blasts (i.e., immature and undifferentiated cells) have a proliferative advantage over normal tissue, proliferating uncontrollably. The cells most involved in this process are the white blood cells, also named lymphocytes, which are produced in large quantities by the bone marrow. These leukemia cells could interrupt their maturation process early and then become resistant to programmed death mechanisms, in this way more cells are produced than they die and these, accumulating in the bone marrow, determine an alteration of proliferation and differentiation of normal hematopoietic cells (e.g., red blood cells and platelets).

There exist different types of Leukemia. They are commonly divided into acute and chronic, based on the rate of progression of the disease. In acute leukemia, the number of cancer cells increases rapidly and the onset of symptoms is early, while in chronic leukemia the malignant cells tend to proliferate more slowly. Over time, however, chronic forms also become aggressive and cause an increase in leukemia cells in the bloodstream. If the disease arises from the lymphoid cells of the bone marrow (from which white blood cells called lymphocytes develop) it is called lymphoid leukemia (ALL), if instead, the starting cell is of the myeloid type (from which red blood cells, platelets and different white blood cells develop from lymphocytes) we speak of myeloid leukemia (AML).



(a) Acute Lymphoblastic Leukemia (ALL).



(b) Acute Myeloid Leukemia (AML).

Figure 5.17: Acute Lymphoblastic Leukemia (a) and Acute Myeloid Leukemia (b).

5.2.2 Dataset

The analyzed data are extracted from microarray dataset, it consists of a large number of gene expressions. Each expression measures the activity level of genes in a particular tissue, enabling us to compare genes expressed in abnormal cancer tissue with those in normal tissue. DNA microarray analysis is useful for simultaneously studying the expression of thousands of genes, and has been rapidly adopted by the research community for the study of a variety of biological processes. It enables to compare two biological classes to identify the differential expression of genes within them, genes with potential relevance to a wide range of biological processes, including cancer development [182].

We have adopted the standard process of microarray analysis [234] depicted in Fig. 5.18. For our analysis, we used data from the microarray Human Illumina 450k Beadchip dataset. It quantifies DNA methylation by treating DNA with sodium bisulfite. The DNA converted to bisulfite is subjected to an amplification step, followed by fragmentation and hybridization to probes on the microarray. The hybridization is allele-specific with a single-base extension of the probes. After this, an out-tag label (ddNTP) is incorporated for detection. The analysis is performed according to the standard protocol provided by Illumina: the DNA is changed to the EZ DNA Methylation kit (Zymo Research), the Bead chip signals are detected and digitized with an Illumina scanner [54]. Bisulfite deaminates unmethylated cytosine, causing its chemical conversion to uracil upon alkaline desulfonation. By selective conversion of cytosine but not 5mC to uracil, followed by PCR and sequencing of cloned amplicon DNA, BGS accurately detects the presence of 5mC in each region of interest at single-nucleotide resolution. After bisulfite conversion, each probe is whole-genome amplified (WGA) and enzymatically fragmented. During hybridization, the WGA-DNA molecules anneal to locus-specific DNA oligomers linked to individual bead types. The two bead types correspond to each CpG locus, one to the methylated (C) and the other to the unmethylated (T) state. Allele-specific primer annealing is followed by single-base extension using DNP-labeled and Biotin-labeled ddNTPs. Both bead types for the same CpG locus will incorporate the same type of labeled nucleotide, determined by the base preceding the interrogated "C" in the CpG locus, and therefore will be detected in the same color channel[5]. The 99% of RefSeq genes are covered, including those in regions of low CpG island density and at risk for being missed by commonly used capture methods[6]. At the end of the process, the chip is scanned to show the intensities of the unmethylated and methylated bead types. The raw data are analyzed and the fluorescence intensity ratios between the two bead types are calculated. A ratio value of 0 represents a non-methylation of

[5] https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf

[6] https://cancergenome.nih.gov/abouttcga/aboutdata/platformdesign/illuminamethylation450

the locus; a ratio of 1 concern total methylation; a value of 0.5 means that one copy is methylated and the other is not, in the diploid human genome.

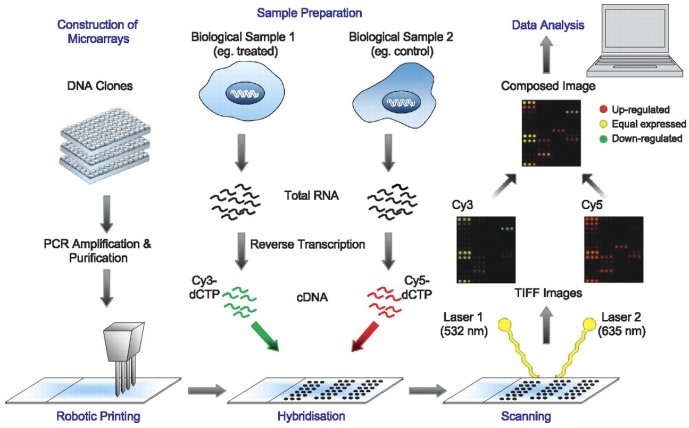


Figure 5.18: Visualization of the process in microarray analysis [234].

The examined dataset was extracted by the GEO platform public database containing genomic data, available on the NCBI website[7]. It consists of biomedical data of 556 patients, where 233 were affected by AML and 323 by ALL. Specifically, for each patient, the dataset contains the detected CpG probes described by their methylation value. These data are related to the Infinium Human Methylation 450k BeadChip microarray, a popular technology to explore DNA methylomes [61].

5.2.3 The proposed approach

To perform feature selection we try to apply and compare two different approaches described in the remaining of this section: (i) Bayesian and autoencoders and (ii) GA.

The Analysis Process based on Bayesian and autoencoders

Analysis of microarray data is based on the hypothesis that the measured fluorescence intensities are representative of the actual level of expression. The complexity of the microarray experimental protocols makes this technology very variable and sometimes subject to significant systematic distortions. For this reason, some manipulations and transformations are necessary before comparing expression levels to attenuate values affected by random aberrations or systematic variations and

[7] https://www.ncbi.nlm.nih.gov/gds

maintain all the data on comparable levels. Figure 5.19 depicts this analysis process consisting of the following six steps.

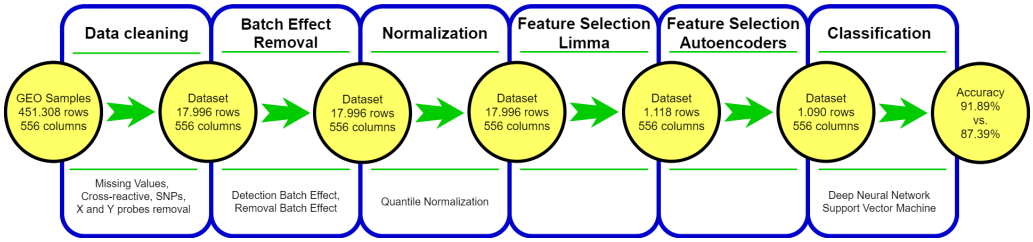


Figure 5.19: The data analysis process based on Bayesian and Autoencoders feature selection.

Step 1 - Data Cleaning The considered samples were extracted from the GEO database and were related to the Illumina Human 450k microarray; they are methylation data related to the GPL13534 platform series. Our initial dataset consisted of 556 samples coming from several microarray experiments. The number of probes among the samples is different; to make the samples uniform and to be able to analyze them we standardized the number of probes among the samples by difference, obtaining 451,308 CpG probes per sample. Then, we performed on our dataset a preprocessing phase by removing the cross-reactive probes, the SNPs probes, and the probes related to sex chromosomes.

Cross-reactive probes target repetitive sequences or co-hybridize alternative sequences that are highly homologous to desired targets and therefore spurious signals can be detected. The cross-reactive sites could reflect CpGs of different methylation status or non-CpGs that are detected as fully methylated or unmethylated loci [43]. Equally important is our search for probes that target CpG sites that overlap with $SNPs$ (single nucleotide polymorphism). SNPs are a variation of the genetic material of a single nucleotide, such that the polymorphic allele is present in the population in a proportion greater than 1%. These portions of the genome can interfere with the methylation analyzes and have to be eliminated. We also remove all the probes related to the X and Y chromosomes because we will focus our analysis only on autosomal genes (not related to sex); this is because there is an imbalance of methylation on sex chromosomes. In particular, the X chromosome, inactive in women, is hypermethylated and this would bring noise into the analysis. In this way, the analysis of genes differentially expressed in the two leukemia types is conducted only on autosomal genes. At this point, we obtained a dataset composed of 556 samples and 434,917 CpG probes. The numerical data within the dataset represent the fluorescent intensities of the probes in double-channel microarray experiments. For the i^{th} probe the estimation of the methylation level $\beta_i \in [0,1]$ is defined

as follows [238]:

$$\beta_i = \frac{max(y_{i,methy}, 0)}{max(y_{i,nmethy}, 0) + max(y_{i,methy}, 0) + \alpha}$$

where y_i is the fluorescent intensity of the probe, $methy$, and $unmethy$ are, respectively, the strength of a methylated and unmethylated signal, and α is an arbitrary value (usually 100) used to stabilize β_i values. On these values we performed a gene sets enrichment analysis operation to obtain the related genes: the resulting dataset was composed of 556 samples and 19,340 genes. A further cleanup eliminated the missing values, as they could generate errors in the measurement and understanding of the relationships between the variables, reducing the genes to 17,996.

Step 2 - Batch Effect Removal The batch effect is a source of variability that has been added to the samples during manipulation, consisting in the introduction of non-biological variability in an experiment [115]. Many factors contribute to the generation of batch effects, some of these include the type of chip, the platform being analyzed, the laboratory, storage conditions, protocols (sample, amplification, labeling and hybridization), cRNA/cDNA synthesis and conditions of washing. In any case, the batch effects often seriously influence the large-scale automatic processing of genomic data sets.

In this step, the batch effect is identified and removed. First, we identified and evaluated the Genomic Spatial Event (GSE) batch variables, i.e., the type of experiment to which each sample refers, through the correlation between the variables. Batch GSEs were identified by statistical analysis of batch medians. This method compares the distribution of each GSE in a single lot to its distribution in all the other lots using the *Kolmogorov-Smirnov (KS) non-parametric test* that verifies the form of the distributions [4]. The p-values returned by the KS test have been corrected by the False Discovery Rate (FDR). This method considers only the biologically relevant differences in the methylation levels through the absolute difference between the median of all the β_i values within a lot for a specific GSE and the respective median of the same GSE in all the other lots. GSEs that had a p-value of significance corrected for FDR lower than 0.01 and had a median difference greater than 0.05 were considered as GSE "batch". After identifying the individual GSE batches, we evaluated the importance of the batch effect in individual batches by considering the number of batch GSEs in the batch and the extent of the deviation of the batch GSE medians in a lot compared to all other lots. We deleted the batch effect on the GSE by using an empirical Bayesian framework and then we validated the results: no further effect was detected.

Step 3 - Normalization The sample data have been normalized to remove systematic variation in a microarray experiment that affects the measured gene expression levels. One of the objectives of DNA microarray analysis is to compare the levels of gene expression in two or more pathological conditions to identify their peculiari-

ties. For our dataset, we have adopted quantile normalization, whose aim is to make equal the empirical intensity distributions of all arrays.

The quantile normalization transforms the intensity distributions of each specific array. In particular, it assigns to each intensity the same value to the quantile to which it belongs. Thus, each intensity has the same distribution in all arrays. This method is based on the consideration that a quantile-quantile graph is a line perfectly coinciding with the diagonal if and only if the distribution of the two data vectors is the same [23].

This means that it is possible to give all arrays the same distribution by replacing the values of the original dataset with the average quantile by applying the following normalization algorithm.

Let M be a matrix of ng genes (rows) and n arrays (columns) representing the number of patients:

1. Sort each column of M, obtaining M_{sort};
2. to each element of the k_{th} row in X_{sort} assign the average value of that row, obtaining M'_{sort};
3. calculate $M_{normalized}$ by reordering each column of M'_{sort} according to the original order.

A negative aspect of this method is that it forces the quantile values to be all the same. This could be a problem in the distribution queues, where it is possible for a gene to have the same value on all arrays, even if this situation rarely occurs [23].

Step 4 - Bayesian Feature Selection In this step, we identified the differentially expressed genes between the two pathological classes AML and ALL by using a Bayesian feature selection technique. Bayesian methods are suitable to study multidimensional inference problems, so it naturally applies to microarray data [77]. Unlike the methods that apply classical inference separately for each gene, the Bayesian analysis exploits information sharing between the genes. We adopted Limma (Linear Models for MicroArray data)[8] to identify these differentially expressed genes through the use of the empirical Bayesian method.

In the following, we describe the adopted feature selection procedure.

Two matrices are obtained from the dataset: the design matrix, containing the samples in the array, and the contrast matrix, which specifies the comparisons to be performed on the samples. The design matrix is specified as follows:

1. the rows represent samples;
2. the columns represent groups. In our case, there exist two columns representing ALL and AML samples, respectively;

[8] https://bioconductor.org/packages/release/bioc/html/limma.html

3. for each sample, the column corresponding to its group has a coefficient equal to 1, otherwise 0.

The contrast matrix represents the difference between the columns.

A first fitting function is applied to the design matrix and data from an experiment involving a series of microarrays with the same set of probes. The function adopts multiple linear models for weighted or generalized minimum squares. Thus, the linear model is adapted to the expression data (by gene) for each probe. A second fitting function is applied to the output of the first regression and the contrast matrix. Given the linear model previously computed, the fitting function calculates the estimated coefficients and standard errors for a given set of contrasts. The function re-orientates the adapted model from the original matrix design to any set of contrasts of the original coefficients. We then applied an empirical Bayesian model for linear data regression, which dynamically borrows information between genes. This function is used to classify genes in order of evidence based on their differential expression; the fact that the same linear model is adapted to each gene allows us to borrow the relationships between genes to moderate residual variances.

Bayesian inference is an approach to statistical inference in which probabilities are not interpreted as frequencies, proportions or similar concepts, but rather as levels of confidence in the occurrence of a given event. The Bayesian model has been very successfully applied in gene expression analyses to moderate the variance estimators gene-wise, and furthermore, in Limma the estimate of global variance can incorporate a tendency to average variance.

We get an estimate of the prior distribution from the marginal distribution of the observed data [179].

The degrees of freedom for the individual variances have increased to reflect the extra information obtained from Bayes' empirical moderation, resulting in an increase in statistical power to detect differential expression [194].

A table of top-ranked genes from the adapted linear model is extracted. This table contains various summary statistics for the top-ranked genes and the selected contrast. In particular, it contains the variables *logFC* and *adj.P.Val*, adopted to perform the feature selection. *logFC* provides the value of the contrast; usually, this represents a change of log_2 times between two or more experimental conditions, while *adj.P.Val* is the distribution of *p-values* adjusted by Benjamini-Hochberg correction [19], which introduced FDR, i.e., the expected proportion of the number of false-positive results on the total of all the positive results, and represents the number of null hypotheses wrongly refused on the total of those rejected. For this reason, we initially extracted the genes with a value of $adj.P.Val < 0.01$ and with a $|logFC| > 2$, thus obtaining 1,118 genes for 556 samples.

Step 5 - Feature Selection with Autoencoders To further reduce our features we used autoencoders. An autoencoder is an unsupervised artificial feed-forward neural network. Conceptually, it is similar to PCA and can be used to reduce the

dimensional space.

The autoencoders compress the input data by forcing the network to use a low-dimensional representation of data. They can to reconstruct the original input. An autoencoder consists of 3 layers: an input layer, a hidden layer and an output layer. The number n of input nodes of this type of network is equal to the number of output nodes (the number of genes). In our case, the hidden layer is composed of two nodes as shown in Fig. 5.20. An autoencoder is divided into two parts: the encoder that learns the mapping between the unlabeled high-dimensional $I[1 : n]$ input data and the low-dimensional representations (in the bottleneck layer), and the decoder that learns the mapping from the intermediate layer representation to the output reconstructed in a high dimension $O[1 : n]$. Autoencoder-based approaches learn to reconstruct input samples by optimizing the Root Mean Squared Error (RMSE) objective function [181].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(I\left[i\right] - O\left[i\right]\right)^2}{n}} \tag{5.4}$$

For autoencoders, we have chosen the batch size equal to the number of genes in our dataset (1,118 genes), Adam's optimization type, an algorithm for the optimization of the gradient of the first order of the stochastic objective functions, based on adaptive estimates of moments of lower order [125]. As activation functions we used "$tanh$", "$linear$", "$tanh$", respectively for the three layers. At the end of the process we obtained 28 abnormal genes that we removed, obtaining a final dataset composed of 1,090 genes.

Step 6 - Classification For the classification, we experimented two different techniques: a deep neural network (DNN) and support vector machines (SVM), as described in the following.

We adopted the DNN shown in Fig. 5.21. The size of the input is equal to the number of genes. The input layer is composed of 30 units with the "$relu$" activation function. The 4 hidden layers transform the representation of the previous layer into a more abstract form. They are composed of 22, 15, 9, 5 respectively, with "$relu$" activation function. The final classification is performed by the last layer (the output layer), composed of 2 units with the "$softmax$" activation function. To classify the class of patients (i.e., AML or ALL) we adopted the "$categorical\ cross\ entropy$" loss function. To set up the hyperparameters we selected the "$adam$" optimizer. We adopted the k-fold cross-validation, with $k = 5$.

Figure 5.22 shows the average loss and average accuracy results on the training and validation sets (blue line and green line, respectively). For the test set we achieved 0,2499 of loss and 91.89% of accuracy.

We also experimented the use of a support vector machine (SVM) [107] as a classifier. SVM are supervised learning models and a powerful technique for classification and regression, with associated learning algorithms. Given a set of training ex-

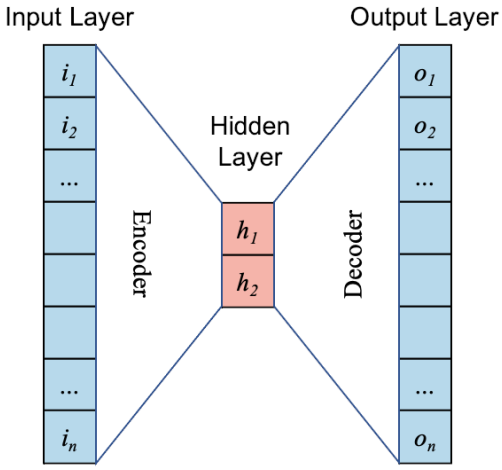


Figure 5.20: The autoencoder architecture, where n represents the number of genes.

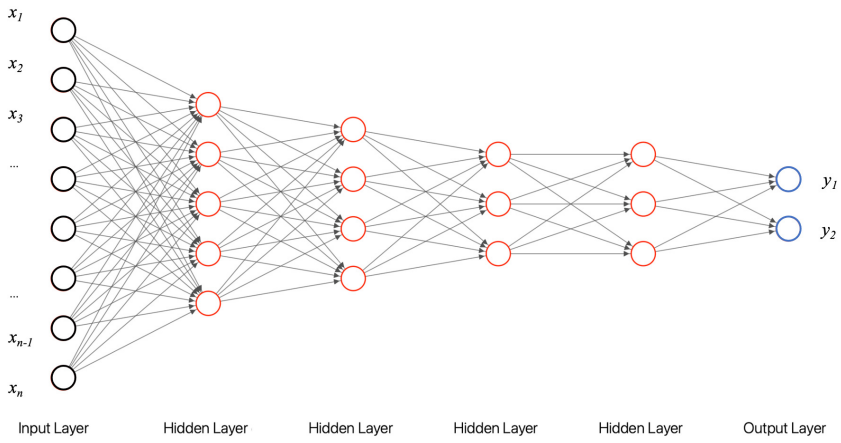


Figure 5.21: The DNN architecture, where n represents the number of genes.

amples, each one belonging to one of two different categories, a training algorithm creates a model that assigns new examples to one or the other category; the model is then used to make predictions for a set of test examples. We got an accuracy of 87,39% by using the same training and test sets adopted with the DNN.

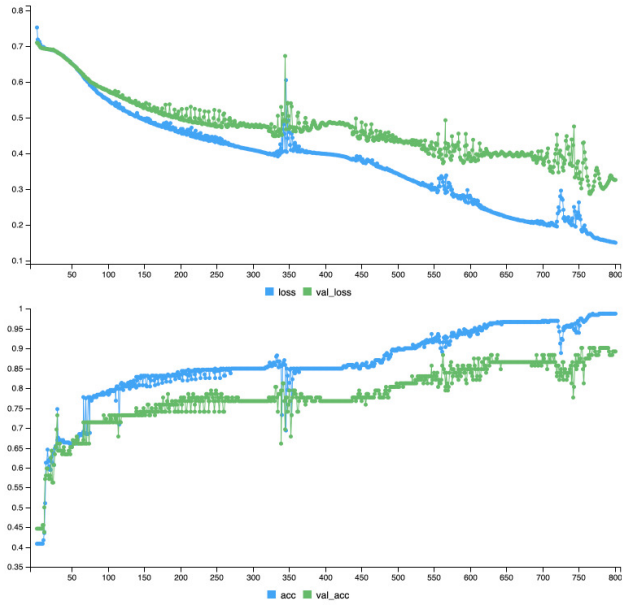


Figure 5.22: Loss and accuracy results of the applied neural network on the feature selection implemented with Limma and autoencoders, where "val"/"$loss$" represents the average accuracy/loss of the training set, and "val_acc"/"val_loss" represents the average accuracy/loss of the validation set, respectively.

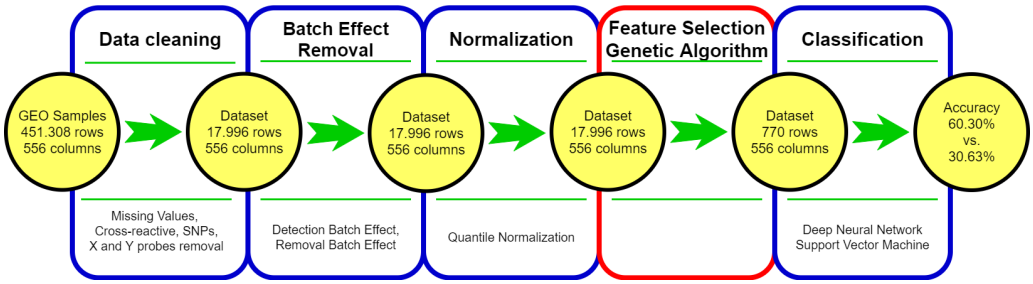


Figure 5.23: The data analysis process based on Genetic Algorithm feature selection.

The analysis process based on a Genetic Algorithm

In this section, we present the second feature selection process we experimented with, based on a genetic algorithm (GA), widely used for this purpose in the litera-

ture (e.g., [17, 198]). In particular, we followed the analysis process shown in Fig. 5.23 that differs from the one in Fig. 5.19 for the red step. Therefore, in the following, we describe only the Feature Selection Genetic Algorithm step and the final classification results.

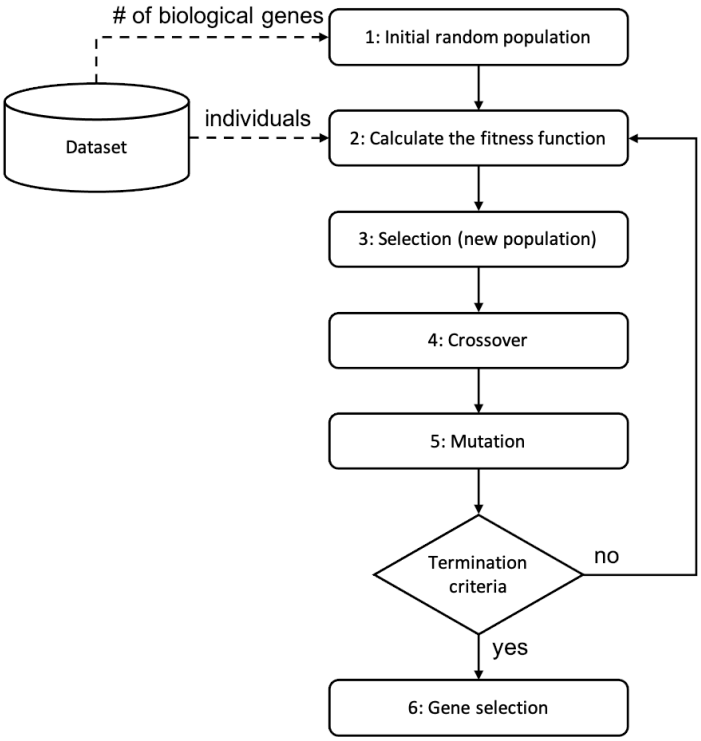


Figure 5.24: Steps of the GA for feature selection.

GAs are heuristic adaptive search algorithms for solving research and optimization problems. They follow a heuristic process (depicted in Fig. 5.24) inspired by the genetics and the principle of natural selection by Charles Darwin. GAs use a set of solutions that evolves at intervals called generations. The evolution is guided by the search for the optimal solution using comparison. In particular, a fitness function is adopted for selecting the best individuals of the current generation that will be used to create the next generation [199]. GA involves a cyclical operation that simulates the evolutionary process of a population. Each cycle represents a generation and consists of operations carried out to generate a new population made up of increasingly better individuals.

Table 5.7: GA parameters.

Parameter	**Value**
Population size	100
Number of generations	100
Fold for cross validation	5
Crossover probability	0.8
Probability of mutation	0.1
Independent probability of crossover	0.8
Independent probability of mutation	0.08
Tournament size	3

Table 5.8: Comparison with the adopted features selection approaches.

Dataset (rows × columns)	**Approach**	**Obtained dataset (rows × columns)**	**Classifier**	**Result accuracy**
17.996×556	Feature Selection Limma + autoencoders	1.090×556	DNN	91.86%
			SVM	87,39%
17.996×556	Feature Selection GA	770×556	DNN	60,36%
			SVM	30,63%

Our dataset consists of 16,408 individuals (i.e., genes). We adopted a GA for feature selection starting from a binary array that represents a chromosome, where genes are the array elements. A chromosome is generally encoded with a bit or character vector. In our case, each element is set to 1 if the biological gene is not expressed (i.e., its value is less equal than 0,5), 0 otherwise. The population is a set of solutions (chromosomes) related to the considered problem.

In the following, we describe the adopted GA, instantiated with the parameters in Table 5.7.

The first step of this algorithm creates the initial population (i.e., 100 individuals) randomly setting the binary values of genes, while the next phases are repeated with each generation and are associated with the principle of natural selection or genetics.

In GAs, individuals have also named chromosomes because of their structure and operations defined on them. Each solution is described by a set of characteristics very similar to the genes and new solutions are created by applying the same mutation and crossover operators present in genetics [208]. The selection of the best individuals is performed by combining or modifying the characteristics that identify an individual. From genetics, the new chromosomes are obtained by recombining their genetic heritage or by changing the genes with the mutation and crossover operators. For each combination of genes, it is possible to calculate a value called fitness which indicates the ability with which the chromosome or solution can solve

the problem. In natural selection, this value measures the individual's adaptation to the environment. So, a better fitness is linked to a greater probability of survival, while within the genetic algorithm to a greater probability of selection. Genetic algorithms are stochastic algorithms in which randomness plays an essential role: both phases of selection and reproduction need procedures involving randomness [157]. Concerning the third step (selection), which favors the selection of better individuals of a population to influence the next generation, we adopted a tournament mechanism. A small number of individuals (i.e., *tournament size* = *3*) is choose randomly with replacement. We keep the fittest one. This is done again and again until you have got 100 individuals.

The fitness function resolves an optimization problem by maximizing the cross-validation accuracy score with the minimum number of selected genes. The score is the accuracy of training data using only the values of the selected genes. In particular, we divided every dataset into 5 equal parts to calculate the fitness value. Then, we selected one of the mentioned parts as a test set and the rest as a training set. We repeated this action five times for every separate part.

In the Crossover phase, the generation of offspring occurs starting from the parents previously selected in the selection phase. The Crossover operator randomly selects a pair of individuals from the pool of solutions for reproduction, with crossover probability (i.e., 0.8); the values of the two solutions are exchanged to generate two new solutions (i.e., the offspring). Crossover aims to generate two new solutions starting from the combination of two previous ones. The crossover probability determines if crossover will happen. A randomly generated floating-point value is compared to this probability, and crossover is performed if the value is less than that probability; otherwise, the offspring is identical to the parents. Moreover, the independent Crossover probability (set to 0,8) concerns the possibility to select a specific gene to perform the exchange between two parents.

The Mutation phase creates a new population starting from the solutions identified in the previous step. Mutation aims to prevent the locking up at a local minimum and to explore the entire research space when each individual in the population reaches a level of fitness close to the average. It mainly maintains genetic diversity within the population. This operator randomly flips (i.e., one becomes zero or vice-versa) some elements of the offspring. Like crossover, there is a mutation probability (set to 0,1). If a randomly selected floating-point value is less than the mutation probability, the mutation is performed on the offspring; otherwise, no mutation occurs. The mutation is performed by randomly selecting (with an independent mutation probability set to 0,08) a gene in the offspring's chromosome and generating a new value uncorrelated to the previous one.

The GA algorithms repeat the process from step 2 until the number of generations (i.e., 100) is reached.

As output, we tried to obtain a better set of individuals that, with the advance-

ment of generations, contains the subsets of genes involved in both AML and ALL. In particular, we selected the genes set to 1. We obtained a reduced dataset containing 6,240 features (genes).

It is worth mentioning the limitations of genetic algorithms. Like most stochastic methods, they do not guarantee success in finding the overall optimal solution to a problem but are often "acceptable" solutions. GAs differ from traditional optimization techniques for several reasons. One of them is that traditional algorithms perform the search starting from a single point while genetic algorithms operate on an entire population of points, and therefore of solutions. This helps in terms of algorithm robustness as it increases the chances of reaching the global optimum and reduces the chances of getting stuck at certain points. For the classification, we used the DNN and the SVM as we did previously for the first process. By applying the DNN to the GA results we obtained 60,36% in accuracy and loss of 0,671. While applying the SVM on the same data we obtained an accuracy of 19,96%.

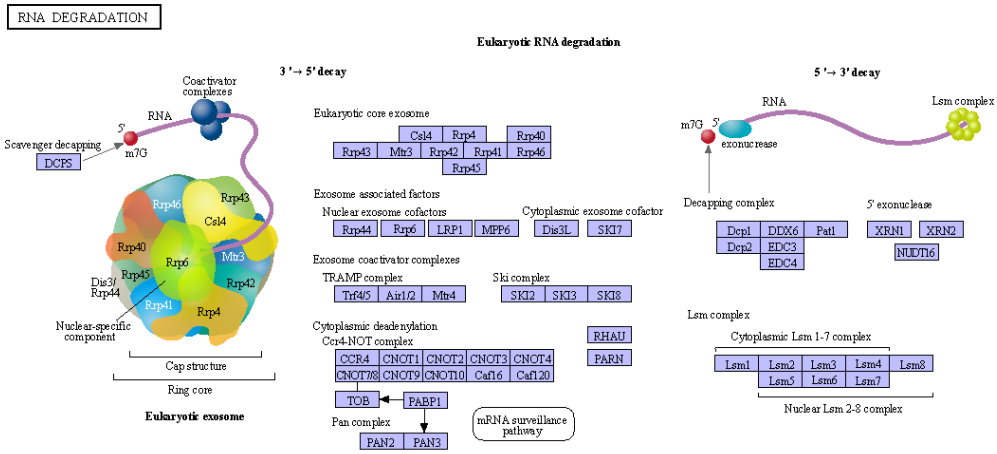


Figure 5.25: The "RNA degradation" pathway [216].

The obtained classification results are summarized in Table 5.8, where we reported for each adopted feature selection process (Limma and the autoencoders vs. GA), the number of selected genes and the classification results obtained with both DNN and SVM classifiers. Results show that feature selection using Limma and the autoencoders performs better with both the classifiers, but the deep neural network (DNN) reached higher accuracy (91,86%) than SVM (87,39%).

Gene	Term	Ont	N	DE	P.DE	FDR
GO:0031981	nuclear lumen	CC	4398	369	4,52E+05	0.000102900918076518
GO:0006396	RNA processing	BP	1270	126	1,47E+05	0.000112329508350203
GO:0044428	nuclear part	CC	4775	392	1,48E+06	0.000112329508350203
GO:0003723	RNA binding	MF	1829	174	2,88E+06	0.000164068843416296
GO:1990904	ribonucleoprotein comp	CC	1281	122	5,62E+06	0.000256002605682553
GO:0031974	membrane-enclosed lum	CC	5518	435	1,63E+07	0.000463503966653521
GO:0043233	organelle lumen	CC	5518	435	1,63E+07	0.000463503966653521
GO:0070013	intracellular organelle l	CC	5518	435	1,63E+07	0.000463503966653521
GO:0016071	mRNA metabolic process	BP	769	88	2,52E+07	0.000638125033014481
GO:0005634	nucleus	CC	7483	563	3,02E+07	0.000671140926002787
GO:0005654	nucleoplasm	CC	3454	297	3,24E+06	0.000671140926002787
GO:0044446	intracellular organelle p	CC	9351	683	1,36E+08	0.00258059896926543
GO:0007052	mitotic spindle organiza	BP	108	22	2,85E+08	0.00498390756368773
GO:0043232	intracellular non-membr	CC	4494	352	5,38E+08	0.00875223356660313
GO:0043228	non-membrane-bounded	CC	4502	352	6,46E+07	0.00980284149640299
GO:0032806	carboxy-terminal domai	CC	19	8	1,15E+09	0.0162912105754246
GO:0044422	organelle part	CC	9640	694	1,47E+09	0.0197241016048166
GO:1902850	microtubule cytoskeleto	BP	129	23	1,71E+09	0.0216675833347165
GO:0005675	transcription factor TFII	CC	11	6	2,05E+09	0.0245830773937175
GO:0043231	intracellular membrane-	CC	10904	773	2,34E+09	0.0254125371642071

Figure 5.26: Gene enriched from "RNA degradation" pathway.

Applying pathways analysis

Pathways analysis provides a means to map key biological processes into important clinical features in disease [71]. It is mainly adopted for predicting cancer outcomes through genome-wide characterizations. In this section, we describe the pathway analysis conducted on the results of the feature selection process based on Limma and autoencoders that reached the best accuracy results and reduced the gene numbers from 19,340 to 1,090. For confirming the relevance at biological level of these results we performed the Gene Sets Enrichment Analysis (GSEA) for the interpretation of gene expression data, which highlights groups of genes that share biological functions (i.e., pathway), chromosomal position or common regulation [213]. In this analysis, we referred to the Kyoto Encyclopedia of Genomes (KEGG)[9], one of the most used databases for pathway knowledge. It links genomic information with functional information of a higher order, computerizing current knowledge on cellular processes and standardizing the genetic annotations [117]. The procedure[10] takes a character vector of significant CpG sites, maps the CpG sites to Entrez Gene IDs to test for GO or KEGG pathway enriched using a hypergeometric test, taking into account the number of CpG sites for gene on EPIC array. In particular, statistical approaches to identify significantly overexpressed CpG groups, by examining p-value and FDR are used. Finally, we have extracted only the pathways with $FDR < 0.05$ [20]. As a result, the analysis detected the "RNA degradation" pathway (see Fig. 5.25) from 20 genes of the 1,090 differentially methylated genes individuated by the feature selection process based on autoencoders. Fig. 5.26 lists the

[9] https://www.genome.jp/kegg

[10] https://bioconductor.org/packages/release/bioc/html/missMethyl.html

pathway's detected genes.

RNA degradation in eukaryotic cells plays a very important role in gene expression, as it balances the transcription rate and also serves to rapidly eliminate transcriptions that are no longer needed. Furthermore, RNA degradation plays a controlling role by eliminating RNA molecules that are considered non-functional or abnormal if they lack sequences or characteristic changes necessary for their functions.

The detected pathway denotes the deregulation of transcription, which is an important factor in the development of leukemia. In particular, in T-cell acute lymphoblastic leukemia (T-ALL) it identifies mutations in the RNA decay factors, including mutations in the CNOT3 gene, which is part of the CCR4-NOT complex that regulates gene expression transcriptionally and post-transcriptionally [46]. This gene is included in the 1,090 we detected and that seems to be involved in mRNA deadenylation. When errors occur in this process, there are quality control mechanisms that detect and eliminate defective transcripts that can lead to dysfunctional or toxic protein. However, these mechanisms do not only ensure the fidelity of RNA transcripts but also perform important regulatory tasks by allowing rapid modulation of steady-state RNA levels in response to changes in the intracellular or extracellular environment [232]. However, it remains unclear how mutations in RNA processing may contribute to the development of leukemia [47].

Chapter 6

DISCUSSION AND CONCLUSION

Artificial intelligence can be a powerful tool in the service of modern medicine. And in part, it already is, as also shown by this thesis work. The applications, in fact, multiply in all areas, from diagnostics to surgery, from drug development to rehabilitation, and are destined to grow.

In this thesis, I presented the results of a long-term study aiming at assessing how Artificial Intelligence may support biomedical data analysis, the main research question **RQ** I proposed. To answer RQ I have concentrated my attention on the identification of degenerative disease, such as Parkinson and Coloboma, and on the experimentation of Artificial Intelligence models and technique in the case of the detection of oncological disease, and I proposed a sub-research question for each of this sub-themes, **RQ1** and **RQ2**, respectively.

Commenting RQ1. Concerning the Parkinson disease, it is difficult to formulate a clinical diagnosis because there are neither objective tests nor specific biochemical and neuroradiological markers. I decided to investigate whether the use of patient records may be useful to classify Parkinson patients. Thus, I applied IR techniques to patient records belonging to a standard dataset for classifying these patients on the base of the reports produced during the different visits. The obtained results are very promising on the use of this technique in the clinical practice. Then, I investigated how the use of traditional biometric techniques may fail in presence of a iris pathology such as Coloboma. Thus, I adopted Artificial Intelligence techniques for detecting irises affected by Coloboma, I demonstrated that traditional biometrics alghorithms fail in presence of this disease, such as the ones proposed by Daugman and Canny. Thus, I developed an algorithm based on image processing which allows also to the people affected by this disease to be recognized by biometrical sys-

tems, and I conducted a preliminary evaluation on a synthetic dataset. Thanks to these results, biometrical systems in the future will respect also for people affected by Coloboma, by assuring also to them they are not excluded by the access a services secured by iris detection, following the General Data Protection Regulation (GDPR) directions.

Commenting RQ2. I conducted researches for experimenting Artificial Intelligence models and technique in the case of the detection of oncological disease. In particular, I concentrated my interests on Leukemia and Melanoma. In the former case I considered two types of Leukemia, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). I proposed a process aiming at detecting a set of differentially expressed genes in terms of methylation level, i.e., genes that in different conditions have an expression level significantly different in the AML and ALL cases, and their characteristic pathways. The detection of gene expression data samples involves feature selection and classification. To this aim, Deep Learning models have been adopted (e.g., feature selection techniques and classifiers methods). A methodology is also proposed for the classification of melanoma by adopting different Deep Learning techniques applied to a common image dataset extracted from the ISIC dataset and consisting of different types of skin diseases, including melanoma on which is applied a specific pre-processing phase. The results of the adopted techniques (i.e., ResNet, 2D CNN, and SOM) are compared to select the best effective neural network for the recognition and classification of melanoma and evaluate the impact of the pre-processing phase. I also propose an augmented reality application for supporting the recognition of the skin lesion in real time, by exploiting both AI and image processing techniques. I describe in detail the real-time process proposed to display the augmented nevus information and evaluated the real-time performances and the app usability. The main results of the proposed approaches are encouraging and suggest that they may be considered in the practical clinical.

In the future, I plan to use Artificial Intelligence models and techniques and image processing to analyze Magnetic Resonance images of the brain to detect progression of Parkinson's disease. Furthermore, the results obtained in the case of Coloboma could be extended and studied in the case of other ophthalmic diseases. Moreover, the augmented reality application for supporting melanoma should experimented in the clinical practice on larger scale.

List of publications

6.0.1 International Journals

- **J01** - Frasca, Maria, and Genoveffa Tortora. "Visualizing correlations among Parkinson biomedical data through information retrieval and machine learning techniques." Multimedia Tools and Applications (2021): 1-19.
- **J02** - Francese, R., Frasca, M., Risi, M. (2021). Are IoBT services accessible to everyone?. Pattern Recognition Letters, 147, 71-77.
- **J03** - Francese, R., Frasca, M., Risi, M., Tortora, G. (2021). A mobile augmented reality application for supporting real-time skin lesion analysis based on deep learning. Journal of Real-Time Image Processing, 1-13 Recognition Letters, 147, 71-77.
- **J04** - Francese, R., Frasca, M., Risi, M.. A deep learning and genetic algorithm based feature selection processes on Leukemia Data. Briefings in Bioinformatics. (Submitted)

6.0.2 International Conferences

- **C01** - Maria Teresa Pellecchia, Maria Frasca, Alessia Auriemma Citarella, Michele Risi, Rita Francese, Genoveffa Tortora, Fabiola De Marco: Identifying Correlations among Biomedical Data through Information Retrieval Techniques. 23rd International Conference on Information Visualization (IV) 2019: 269-274.
- **C02** - Francese, R., Frasca, M., Guarino, A., Malandrino, D., Risi, M., Zaccagnino, R., Lettieri, N. (2020, September). On the Limitation of Pathological Iris Recognition: Neural Network Perspectives. In 2020 24th International Conference Information Visualisation (IV) (pp. 68-73). IEEE.
- **C03** - Frasca, M., Nappi, M., Risi, M., Tortora, G., Citarella, A. A. (2021, January). A comparison of neural network approaches for melanoma classifica-

tion. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 2110-2117). IEEE.

- **C04** - Francese, R., Frasca, M., Risi, M., Tortora, G. (2020, September). An augmented reality mobile application for skin lesion data visualization. In 2020 24th International Conference Information Visualisation (IV) (pp. 51-56). IEEE.

Bibliography

[1] N. R. Abbasi, H. M. Shaw, D. S. Rigel, R. J. Friedman, W. H. McCarthy, I. Osman, A. W. Kopf, and D. Polsky. Early diagnosis of cutaneous melanoma: revisiting the abcd criteria. *Jama*, 292(22):2771–2776, 2004.

[2] M. Abdelsamea, M. H. Mohamed, and M. Bamatraf. An effective image feature classiffication using an improved som. *arXiv preprint arXiv:1501.01723*, 2015.

[3] O. Abuzaghleh, B. D. Barkana, and M. Faezipour. Noninvasive real-time automated skin lesion analysis system for melanoma early detection and prevention. *IEEE Journal of Translational Eng. in Health and Medicine*, 3:1–12, 2015.

[4] R. Akulenko, M. Merl, and V. Helms. BEclear: batch effect detection and adjustment in DNA methylation data. *PloS One*, 11(8):1–17, 2016.

[5] M. S. Al-Ahwal. Chemotherapy and fingerprint loss: beyond cosmetic. *The oncologist*, 17(2):291, 2012.

[6] K. Alsabti, S. Ranka, and V. Singh. An efficient k-means clustering algorithm. 1997.

[7] A. M. Alsamman, S. D. Ibrahim, and A. Hamwieh. Kaspspoon: an in vitro and in silico pcr analysis tool for high-throughput snp genotyping. *Bioinformatics*, 35(17):3187–3190, 2019.

[8] A. Anagaw and Y.-L. Chang. A new complement naïve bayesian approach for biomedical data classification. *Journal of Ambient Intelligence and Humanized Computing*, 10(10):3889–3897, 2019.

[9] G. Argenziano, C. Catricalà, M. Ardigo, P. Buccini, P. De Simone, L. Eibenschutz, A. Ferrari, G. Mariani, V. Silipo, I. Sperduti, et al. Seven-point checklist of dermoscopy revisited. *British Journal of Dermatology*, 164(4):785–790, 2011.

[10] K. Ashique, F. Kaliyadan, and S. J. Aurangabadkar. Clinical photography in dermatology using smartphones: An overview. *Indian Dermatology Online Journal*, 6(3):158, 2015.

[11] T. M. Aslam, S. Z. Tan, and B. Dhillon. Iris recognition in the presence of ocular disease. *Journal of The Royal Society Interface*, 6(34):489–493, 2009.

[12] T. K. Attwood, M. D. Croning, D. R. Flower, A. Lewis, J. Mabey, P. Scordis, J. Selley, and W. Wright. Prints-s: the database formerly known as prints. *Nucleic Acids Research*, 28(1):225–227, 2000.

[13] C. M. Balch, A. C. Buzaid, S.-J. Soong, M. B. Atkins, N. Cascinelli, D. G. Coit, I. D. Fleming, J. E. Gershenwald, A. Houghton Jr, J. M. Kirkwood, et al. Final version of the american joint committee on cancer staging system for cutaneous melanoma. *Journal of Clinical Oncology*, 19(16):3635–3648, 2001.

[14] M. Bär, P. Tschandl, and H. Kittler. Differentiation of pigmented spitz nevi and reed nevi by integration of dermatopathologic and dermatoscopic findings. *Derm. Practical & Conceptual*, 2(1):13, 2012.

[15] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques. Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, 8(3):965–979, 2013.

[16] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, 2007.

[17] B. Baur and S. Bozdag. A feature selection algorithm to compute gene centric methylation from probe level methylation data. *PloS One*, 11(2):1–19, 2016.

[18] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. S. Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486*, 2018.

[19] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300, 1995.

[20] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[21] J. Blaas, C. P. Botha, and F. H. Post. Interactive visualization of multi-field medical data using linked physical and feature-space views. In *EuroVis*, pages 123–130, 2007.

[22] S. Bleik, M. Mishra, J. Huan, and M. Song. Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(5):1211–1217, 2013.

[23] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[24] M. R. Bouadjenek and K. Verspoor. Multi-field query expansion is effective for biomedical dataset retrieval. *Database*, 2017, 2017.

[25] P. E. Bourne. Bioinformatics meets data mining: time to dance? *Trends in biotechnology*, 18(6):228–230, 2000.

[26] A. Boyd, A. Czajka, and K. Bowyer. Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch? *arXiv preprint arXiv:2002.08916*, 2020.

[27] P. Boyle and B. Levin. Worldwide cancer burden. *World cancer report*, pages 16–54, 2014.

[28] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019.

[29] J. Brooke. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[30] T. Burchardt, J. Le Grand, D. Piachaud, J. Hills, and L. Grand. Understanding social exclusion, 2002.

[31] R. J. Burton, M. Albur, M. Eberl, and S. M. Cuff. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC medical informatics and decision making*, 19(1):1–11, 2019.

[32] K. Cahyaningrum, W. Astuti, et al. Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence. In *Procs. of the International Conference on Data Science and Its Applications (ICoDSA)*, pages 1–7. IEEE, 2020.

[33] P. Campisi. *Security and privacy in biometrics*, volume 24. Springer, 2013.

[34] P. Carli, V. De Giorgi, E. Crocetti, L. Caldini, C. Ressel, and B. Giannotti. Diagnostic and referral accuracy of family doctors in melanoma screening: effect of a short formal training. *European Journal of Cancer Prevention*, 14(1):51–55, 2005.

[35] P. B. Castro, B. Krohling, A. G. Pacheco, and R. A. Krohling. An app to detect melanoma using deep learning: An approach to handle imbalanced data based on evolutionary algorithms. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2020.

[36] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss. A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362–373, 2007.

[37] R. Chaturvedi and Y. Thakur. Iris recognition using Daugmanâs algorithm and ann. *Intl. Journal of Applied Engineering Research*, 14(21):3987–3995, 2019.

[38] B. B. Chaudhuri and N. Sarkar. Texture segmentation using fractal dimension. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):72–77, 1995.

[39] H. Chen, S. S. Fuller, C. Friedman, and W. Hersh. Knowledge management, data mining, and text mining in medical informatics. In *Medical Informatics*, pages 3–33. Springer, 2005.

[40] H. Chen, Y. Zhang, and I. Gutman. A kernel-based clustering method for gene selection with gene expression data. *Journal of Biomedical Informatics*, 62:12–20, 2016.

[41] Q. Chen and M. Sokolova. Word2vec and doc2vec in unsupervised sentiment analysis of clinical discharge summaries. *arXiv preprint arXiv:1805.00352*, 2018.

[42] R. H. Chen, M. Snorrason, S. M. Enger, E. Mostafa, J. M. Ko, V. Aoki, and J. Bowling. Validation of a skin-lesion image-matching algorithm based on computer vision technology. *Telemedicine and e-Health*, 22(1):45–50, 2016.

[43] Y.-a. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2):203–209, 2013.

[44] B.-J. Cho, Y. J. Choi, M.-J. Lee, J. H. Kim, G.-H. Son, S.-H. Park, H.-B. Kim, Y.-J. Joo, H.-Y. Cho, M. S. Kyung, et al. Classification of cervical neoplasms on

colposcopic photography using deep learning. *Scientific reports*, 10(1):1–10, 2020.

[45] S. Chou, W. Chang, C.-Y. Cheng, J.-C. Jehng, and C. Chang. An information retrieval system for medical records & documents. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1474–1477. IEEE, 2008.

[46] M. A. Collart and O. O. Panasenko. The Ccr4–not complex. *Gene*, 492(1):42–53, 2012.

[47] J. Cools. ART: aberrant RNA degradation in T-cell leukemia. `https://cordis.europa.eu/project/rcn/185655/factsheet`, 2014-2019. [Online].

[48] M. D. Corbo and J. Wismer. Agreement between dermatologists and primary care practitioners in the diagnosis of malignant melanoma: review of the literature. *Journal of Cutaneous Medicine and Surgery*, 16(5):306–310, 2012.

[49] F. Cozza, A. Guarino, F. Isernia, D. Malandrino, A. Rapuano, R. Schiavone, and R. Zaccagnino. Hybrid and lightweight detection of third party tracking: Design, implementation, and evaluation. *Computer Networks*, 167:106993, 2020.

[50] M. Craven, J. Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86, 1999.

[51] A. Criminisi, P. Perez, and K. Toyama. Object removal by exemplar-based inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–II. IEEE, 2003.

[52] V. Dal Pozzo, C. Benelli, and E. Roscetti. The seven features for melanoma: a new dermoscopic algorithm for the diagnosis of malignant melanma. *European Journal of Dermatology*, 9(4):303–8, 1999.

[53] M. Daoud and M. Mayo. A survey of neural network-based cancer prediction models from microarray data. *Artificial Intelligence in Medicine*, 97:204–214, 2019.

[54] R. P. Darst, C. E. Pardo, L. Ai, K. D. Brown, and M. P. Kladde. Bisulfite sequencing of DNA. *Current Protocols in Molecular Biology*, pages 7–9, 2010.

[55] J. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, 1993.

[56] J. Daugman. The importance of being random: Statistical principles of iris recognition. *Pattern recognition*, 36(2):279–291, 2003.

[57] J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(5):1167–1175, 2007.

[58] J. Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009.

[59] C. A. Davie. A review of parkinson's disease. *British medical bulletin*, 86(1):109–127, 2008.

[60] R. de Luis-GarcıÌa, C. Alberola-López, O. Aghzout, and J. Ruiz-Alzola. Biometric identification systems. *Signal processing*, 83(12):2539–2557, 2003.

[61] S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics*, 15(6):929–941, 2014.

[62] P. Deloukas, G. Schuler, G. Gyapay, E. Beasley, C. Soderlund, P. Rodriguez-Tome, L. Hui, T. Matise, K. McKusick, J. Beckmann, et al. A physical map of 30,000 human genes. *Science*, 282(5389):744–746, 1998.

[63] X. Deng and Y. Xu. Cancer classification using microarray data by DPCAForest. In *Procs. of the International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1081–1087. IEEE, 2019.

[64] Dermengine. `https://www.dermengine.com`.

[65] H. Dhrif, L. G. S. Giraldo, M. Kubat, and S. Wuchty. A stable combinatorial particle swarm optimization for scalable feature selection in gene expression data. *arXiv preprint arXiv:1901.08619*, 2019.

[66] C. Dirican. The impacts of robotics, artificial intelligence on business and economics. *Procedia-Social and Behavioral Sciences*, 195:564–573, 2015.

[67] D. Distante, M. Risi, and G. Scanniello. Extending web content management systems navigation capabilities with semantic navigation maps. In *12th IEEE Intl. Symposium on Web Systems Evolution (WSE)*, pages 1–5. IEEE, 2010.

[68] T. Do, T. Hoang, V. Pomponiu, Y. Zhou, Z. Chen, N. Cheung, D. Koh, A. Tan, and S. Tan. Accessible melanoma detection using smartphones and mobile image analysis. *IEEE Trans. on Multimedia*, 20(10):2849–2864, 2018.

[69] B. S. dos Santos, M. T. A. Steiner, A. T. Fenerich, and R. H. P. Lima. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering*, 138:106120, 2019.

[70] E. Dynomant, S. J. Darmoni, É. Lejeune, G. Kerdelhué, J.-P. Leroy, V. Lequertier, S. Canu, and J. Grosjean. Doc2vec on the pubmed corpus: study of a new approach to generate related articles. *arXiv preprint arXiv:1911.11698*, 2019.

[71] S. Efroni, C. F. Schaefer, and K. H. Buetow. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PloS One*, 2(5):e425, 2007.

[72] F. El Khoury. *Iris biometric model for secured network access*. CRC Press, 2016.

[73] F. Ercal, A. Chawla, W. V. Stoecker, H.-C. Lee, and R. H. Moss. Neural network diagnosis of malignant melanoma from color images. *IEEE Transactions on Biomedical Engineering*, 41(9):837–845, 1994.

[74] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[75] J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In *IJCAI*, volume 7, pages 348–353, 2007.

[76] E. Fernández, J.-M. García-Moreno, A. Martín de Pablos, and J. Chacón. May the thyroid gland and thyroperoxidase participate in nitrosylation of serum proteins and sporadic parkinson's disease?, 2014.

[77] C. K. Fisher and P. Mehta. Bayesian feature selection for high-dimensional linear regression via the Ising approximation with applications to genomics. *Bioinformatics*, 31(11):1754–1761, 01 2015.

[78] A. Floratos, I. Jurisica, and I. Rigoutsos. Knowledge discovery in biological domains (tutorial am-3). In *Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 95–163, 2000.

[79] R. Francese, M. Frasca, A. Guarino, D. Malandrino, M. Risi, R. Zaccagnino, and N. Lettieri. On the limitation of pathological iris recognition: Neural network perspectives. In *2020 24th International Conference Information Visualisation (IV)*, pages 68–73. IEEE, 2020.

[80] R. Francese, M. Frasca, and M. Risi. Are iobt services accessible to everyone? *Pattern Recognition Letters*, 147:71–77, 2021.

[81] R. Francese, M. Frasca, M. Risi, and G. Tortora. A mobile augmented reality application for supporting real-time skin lesion analysis based on deep learning. *Journal of Real-Time Image Processing*, pages 1–13, 2021.

[82] M. Frasca, M. Nappi, M. Risi, G. Tortora, and A. A. Citarella. A comparison of neural network approaches for melanoma classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2110–2117. IEEE, 2021.

[83] M. Frasca and G. Tortora. Visualizing correlations among parkinson biomedical data through information retrieval and machine learning techniques. *Multimedia Tools and Applications*, pages 1–19, 2021.

[84] K. A. Freedberg, A. C. Geller, D. R. Miller, R. A. Lew, and H. K. Koh. Screening for malignant melanoma: a cost-effectiveness analysis. *Journal of the American Academy of Dermatology*, 41(5):738–745, 1999.

[85] R. J. Friedman, D. Rigel, M. K. Silverman, A. W. Kopf, and K. A. Vossaert. Malignant melanoma in the 1990s: The continued importance of earlydetection and the role of physician examination and self-examination of theskin. *CA: A Cancer Journal for Clinicians*, 41(4):201–226, 1991.

[86] R. J. Friedman, D. S. Rigel, and A. W. Kopf. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians*, 35(3):130–151, 1985.

[87] Z. Gao, J. Zhang, L. Zhou, and L. Wang. Hep-2 cell image classification with convolutional neural networks. In *2014 1st Workshop on Pattern Recognition Techniques for Indirect Immunofluorescence Images*, pages 24–28. Ieee, 2014.

[88] I. Gath and A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, (7):773–780, 1989.

[89] G. Gautam and S. Mukhopadhyay. Contact lens detection using transfer learning with deep representations. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[90] D. Gefen, J. Miller, J. K. Armstrong, F. H. Cornelius, N. Robertson, A. Smith-McLallen, and J. A. Taylor. Identifying patterns in medical records through latent semantic analysis. *Communications of the ACM*, 61(6):72–77, 2018.

[91] D. J. Gelb, E. Oliver, and S. Gilman. Diagnostic criteria for parkinson disease. *Archives of neurology*, 56(1):33–39, 1999.

[92] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik. Recursive memetic algorithm for gene selection in microarray data. *Expert Systems with Applications*, 116:172–185, 2019.

[93] H. Gite and C. Mahender. Iris code generation and recognition. *International Journal of Machine Intelligence*, 3(3):103–107, 2011.

[94] V. Gokul Rajan and S. Vijayalakshmi. A new approach for sclera segmentation using integro differential operator. *Journal of Computational and Theoretical Nanoscience*, 17(5):2330–2335, 2020.

[95] W. N. Grundy and T. L. Bailey. Family pairwise search with embedded motif models. *Bioinformatics (Oxford, England)*, 15(6):463–470, 1999.

[96] A. Guffanti and G. Simon. Uniblast and the est extractor: new www resources for est data mining. *Trends in genetics*, 14(7):293, 1998.

[97] D. Gupta and M. P. K. Sharma. Performance measurement of edge detectors for human iris segmentation and detection. *Intl. Journal of Engineering and Technical Research*, 9(1):58, 2019.

[98] L. Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161, 1954.

[99] M. Haenlein and A. Kaplan. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4):5–14, 2019.

[100] T. Hao, X. Chen, G. Li, and J. Yan. A bibliometric analysis of text mining in medical research. *Soft Computing*, 22(23):7875–7892, 2018.

[101] D. Harman. Ranking algorithms. in information retrieval: Data structures and algorithms. *WB Frakes and R. Baeza-Yates, Eds. Prentice Hall*, pages 363–392, 1992.

[102] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[103] P. Helsing and M. Loeb. Small diameter melanoma: a follow-up of the norwegian melanoma project. *British Journal of Dermatology*, 151(5):1081–1083, 2004.

[104] K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The prosite database, its status in 1999. *Nucleic Acids Research*, 27(1):215–219, 1999.

[105] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[106] M. Hossin and M. Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[107] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al. A practical guide to support vector classification, 2003.

[108] G. Hu. Total cholesterol and the risk of parkinson's disease: a review for some new findings. *Parkinsonâs disease*, 2010, 2010.

[109] S. A. H. Ibrahim. Melanoma detection using mobile technology and feature-based classification techniques. *Thesis*, 2014.

[110] A. Ignatov, R. Timofte, W. Chou, K. Wang, M. Wu, T. Hartley, and L. Van Gool. Ai benchmark: Running deep neural networks on android smartphones. In *Procs. of the European Conf. on Computer Vision (ECCV)*, 2018.

[111] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[112] A. Jahwar and N. Ahmed. Swarm intelligence algorithms in gene selection profile based on classification of microarray data: a review. *Journal of Applied Science and Technology Trends*, 2(01):01–09, 2021.

[113] A. K. Jain, R. Bolle, and S. Pankanti. *Biometrics: personal identification in networked society*, volume 479. Springer Science & Business Media, 2006.

[114] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004.

[115] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[116] M. L. Juhász and E. S. Marmur. Reviewing challenges in the diagnosis and treatment of lentigo maligna and lentigo-maligna melanoma. *Rare Cancers and Therapy*, 3(1-2):133–145, 2015.

[117] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[118] C. Kang, Y. Huo, L. Xin, B. Tian, and B. Yu. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of Theoretical Biology*, 463:77–91, 2019.

[119] N. Karimian, P. A. Wortman, and F. Tehranipoor. Evolving authentication design considerations for the internet of biometric things (IoBT). In *Intl. Conf. on Hardware/Software Codesign and System Synthesis*, pages 1–10, 2016.

[120] A. Kassianos, J. Emery, P. Murchie, and F. M. Walter. Smartphone applications for melanoma detection by community, patient and generalist clinician users: a review. *British J. of Derm.*, 172(6):1507–1518, 2015.

[121] T. Kerikmäe and A. Rull. *The future of law and etechnologies*, volume 3. Springer, 2016.

[122] A. Khan, B. Baharudin, L. H. Lee, and K. Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.

[123] S. Kido, Y. Hirano, and N. Hashimoto. Detection and classification of lung abnormalities by use of convolutional neural network (cnn) and regions with cnn features (r-cnn). In *International Workshop on Advanced Image Technology (IWAIT)*, pages 1–4, 2018.

[124] G. King and L. Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[125] D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[126] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman. 1d convolutional neural networks and applications: A survey. *arXiv preprint arXiv:1905.03554*, 2019.

[127] J. Koh, M. Suk, and S. M. Bhandarkar. A multilayer self-organizing feature map for range image segmentation. *Neural Networks*, 8(1):67–86, 1995.

[128] T. Kothmayr, C. Schmitt, W. Hu, M. Brünig, and G. Carle. A dtls based end-to-end security architecture for the internet of things with two-way authentication. In *37th Annual IEEE Conference on Local Computer Networks-Workshops*, pages 956–963. IEEE, 2012.

[129] M. A. Kowtko. Biometric authentication for older adults. In *IEEE Long Island Systems, Applications and Technology (LISAT) Conference 2014*, pages 1–6. IEEE, 2014.

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

[131] J. Kumuthini, M. Chimenti, S. Nahnsen, A. Peltzer, R. Meraba, R. McFadyen, G. Wells, D. Taylor, M. Maienschein-Cline, J.-L. Li, et al. Ten simple rules for providing effective bioinformatics research support, 2020.

[132] R. Laganiere. A morphological operator for corner detection. *Pattern Recognition*, 31(11):1643–1652, 1998.

[133] K. Lan, D.-t. Wang, S. Fong, L.-s. Liu, K. K. Wong, and N. Dey. A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, 42(8):1–20, 2018.

[134] S. Laotrakunchai, A. Wongkaew, and K. Patanukhom. Measurement of size and distance of objects using mobile devices. In *Intl. Conf. on Signal-Image Technology Internet-Based Systems*, pages 156–161, 2013.

[135] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[136] R. Leenes. Framing techno-regulation: An exploration of state and non-state regulation by technology. *Legisprudence*, 5(2):143–169, 2011.

[137] A. Lesk. *Introduction to bioinformatics*. Oxford university press, 2019.

[138] B. J. Lesselroth and D. S. Pieczkiewicz. *Data visualization strategies for the electronic health record*. Nova Science Publishers, Inc., 2011.

[139] N. Lettieri, A. Guarino, D. Malandrino, and R. Zaccagnino. Platform economy and techno-regulationâexperimenting with reputation and nudge. *Future Internet*, 11(7):163, 2019.

[140] Q. Li and Y.-F. B. Wu. Identifying important concepts from medical documents. *Journal of biomedical informatics*, 39(6):668–679, 2006.

[141] J. Liu, Y. Liu, C. Wang, A. Li, B. Meng, X. Chai, and P. Zuo. An original neural network for pulmonary tuberculosis diagnosis in radiographs. In *International Conference on Artificial Neural Networks*, pages 158–166. Springer, 2018.

[142] D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405(6788):827–836, 2000.

[143] C. Lu, M. Mahmood, N. Jha, and M. Mandal. Automated segmentation of the melanocytes in skin histopathological images. *IEEE Journal of Biomedical and Health Informatics*, 17(2):284–296, 2013.

[144] J. Lv, Q. Peng, X. Chen, and Z. Sun. A multi-objective heuristic algorithm for gene expression microarray data classification. *Expert Systems with Applications*, 59:13–19, 2016.

[145] M. Lyksborg, O. Puonti, M. Agn, and R. Larsen. An ensemble of 2d convolutional neural networks for tumor segmentation. In *Scandinavian Conference on Image Analysis*, pages 201–211. Springer, 2015.

[146] O. H. MacLin and R. S. Malpass. The ambiguous-race face illusion. *Perception*, 32(2):249–252, 2003.

[147] T. Maier, D. Kulichova, K. Schotten, R. Astrid, T. Ruzicka, C. Berking, and A. Udrea. Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. *Journal of the European Academy of Dermatology and Venereology*, 29(4):663–667, 2015.

[148] S. Majumder and M. A. Ullah. Feature extraction from dermoscopy images for an effective diagnosis of melanoma skin cancer. In *10th International Conference on Electrical and Computer Engineering (ICECE)*, pages 185–188. IEEE, 2018.

[149] W. Mao and W. W. Chu. The phrase-based vector space model for automatic retrieval of free-text medical documents. *Data & Knowledge Engineering*, 61(1):76–92, 2007.

[150] S. N. Markovic, L. A. Erickson, R. D. Rao, R. R. McWilliams, L. A. Kottschade, E. T. Creagan, R. H. Weenig, J. L. Hand, M. R. Pittelkow, B. A. Pockaj, et al. Malignant melanoma in the 21st century, part 1: epidemiology, risk factors, screening, prevention, and diagnosis. In *Mayo Clinic Proceedings*, volume 82, pages 364–380. Elsevier, 2007.

[151] M. A. Marra, L. Hillier, and R. H. Waterston. Expressed sequence tagsâestablishing bridges between genomes. *Trends in Genetics*, 14(1):4–7, 1998.

[152] E. Marshall. Do-it-yourself gene watching, 1999.

[153] L. Masek et al. *Recognition of human iris patterns for biometric identification*. PhD thesis, Citeseer, 2003.

[154] H. Menon and A. Mukherjee. Iris biometrics using deep convolutional networks. In *IEEE Intl. Instrumentation and Measurement Technology Conf.*, pages 1–5. IEEE, 2018.

[155] B. Meskó, Z. Drobni, É. Bényei, B. Gergely, and Z. Győrffy. Digital health is a cultural transformation of traditional healthcare. *Mhealth*, 3, 2017.

[156] S. Minaee and A. Abdolrashidi. Deepiris: Iris recognition using a deep learning approach. *arXiv preprint arXiv:1907.09380*, 2019.

[157] S. Mirjalili. Genetic algorithm. In *Evolutionary Algorithms and Neural Networks*, pages 43–55. Springer, 2019.

[158] M. Mohammadi, H. S. Noghabi, G. A. Hodtani, and H. R. Mashhadi. Robust and stable gene selection via maximum–minimum correntropy criterion. *Genomics*, 107(2-3):83–87, 2016.

[159] J. Moor. *The Turing test: the elusive standard of artificial intelligence*, volume 30. Springer Science & Business Media, 2003.

[160] E. Mordini and D. Tzovaras. *Second generation biometrics: The ethical, legal and social context*, volume 11. Springer Science & Business Media, 2012.

[161] C. Morton and R. Mackie. Clinical accuracy of the diagnosis of cutaneous malignant melanoma. *The British Journal of Dermatology*, 138(2):283–287, 1998.

[162] R. P. Munhoz, H. A. Teive, A. R. Troiano, P. R. Hauck, M. H. H. Leiva, H. Graff, and L. C. Werneck. Parkinson's disease and thyroid dysfunction. *Parkinsonism & related disorders*, 10(6):381–383, 2004.

[163] A. Namozov and Y. Im Cho. Convolutional neural network algorithm with parameterized activation function for melanoma classification. In *International Conference on Information and Communication Technology Convergence (ICTC)*, pages 417–419. IEEE, 2018.

[164] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M. R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian. Melanoma detection by analysis of clinical images using convolutional neural network. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1373–1376. IEEE, 2016.

[165] W. Newell, S. Beck, H. Lehrach, and A. Lyall. Estimation of distances and map construction using radiation hybrids. *Genome research*, 8(5):493–508, 1998.

[166] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan. Iris recognition with off-the-shelf cnn features: A deep learning perspective. *IEEE Access*, 6:18848–18855, 2017.

[167] H. Ohmaid, S. Eddarouich, A. Bourouhou, and M. Timouyas. Iris segmentation using a new unsupervised neural approach. *IAES Intl. Journal of Artificial Intelligence*, 9(1):58, 2020.

[168] E. Okur and M. Turkan. A survey on automated melanoma detection. *Engineering Applications of Artificial Intelligence*, 73:50–67, 2018.

[169] M. V. Olson. A time to sequence. *Science*, 270(5235):394–396, 1995.

[170] M. S. Othman, S. R. Kumaran, and L. M. Yusuf. Gene selection using hybrid multi-objective Cuckoo search algorithm with evolutionary operators for cancer microarray data. *IEEE Access*, 8:186348–186361, 2020.

[171] A. G. Pacheco and R. A. Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in biology and medicine*, 116:103545, 2020.

[172] R. A. Pagon. Ocular Coloboma. *Survey of ophthalmology*, 25(4):223–236, 1981.

[173] S. Pai and A. Bhardwaj. Eye gesture based communication for people with motor disabilities in developing nations. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[174] M. T. Pellecchia, M. Frasca, A. A. Citarella, M. Risi, R. Francese, G. Tortora, and F. De Marco. Identifying correlations among biomedical data through information retrieval techniques. In *2019 23Rd international conference information visualisation (IV)*, pages 269–274. IEEE, 2019.

[175] E. Perera, N. Gnaneswaran, R. Jennens, and R. Sinclair. Malignant melanoma. *Healthcare*, 2:1–19, 2013.

[176] I. Petrov and N. Minakova. Optimization method for non-cooperative iris recognition task using Daugman integro-differential operator. In *Journal of Physics: Conference Series*, volume 1615, page 012007. IOP Publishing, 2020.

[177] R. Peverelli and R. D. Feniks. Skinvision: Leading mobile solution to monitor, track and understand skin health. *Digital Insurance Agenda*, 2017. https://www.digitalinsuranceagenda.com/featured-insurtechs/skinvision-leading-mobile-solution-to-monitor-\track-and-understand-skin-health/.

[178] M. Phillips, J. Greenhalgh, H. Marsden, and I. Palamaras. Detection of malignant melanoma using artificial intelligence: An observational study of diagnostic accuracy. *Derm. Practical & Conceptual*, 10(1), 2020.

[179] B. Phipson, S. Lee, I. J. Majewski, W. S. Alexander, and G. K. Smyth. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics*, 10(2):946, 2016.

[180] H. Proença and J. C. Neves. Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Transactions on Information Forensics and Security*, 13(4):888–896, 2017.

[181] A. Pumsirirat and L. Yan. Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of Advanced Computer Science and Applications*, 9(1):18–25, 2018.

[182] J. Quackenbush. Microarray analysis and tumor classification. *New England Journal of Medicine*, 354(23):2463–2472, 2006.

[183] Qualcomm. Qualcomm neural processing sdk for AIDermengine. https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk.

[184] N. H. Quang et al. Automatic skin lesion analysis towards melanoma detection. In *21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)*, pages 106–111. IEEE, 2017.

[185] A. Rajaraman and J. D. Ullman. *Data Mining*, page 1â17. Cambridge University Press, 2011.

[186] A. Ramesh, C. Kambhampati, J. R. Monson, and P. Drew. Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England*, 86(5):334, 2004.

[187] S. Rastan and L. J. Beeley. Functional genomics: going forwards from the databases. *Current opinion in genetics & development*, 7(6):777–783, 1997.

[188] N. Razmjooy, B. S. Mousavi, F. Soleymani, and M. H. Khotbesara. A computer-aided diagnosis system for malignant melanomas. *Neural Computing and Applications*, 23(7-8):2059–2071, 2013.

[189] J. Reed. Trends in commercial bioinformatics. *Oscar Gruss Biotechnology Review*, 13, 2000.

[190] F. M. Riese and S. Keller. Susi: Supervised self-organizing maps for regression and classification in python. *arXiv preprint arXiv:1903.11114*, 2019.

[191] F. M. Riese, S. Keller, and S. Hinz. Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data. *Remote Sensing*, 12(1):7, 2020.

[192] D. S. Rigel, J. Russak, and R. Friedman. The evolution of melanoma diagnosis: 25 years beyond the abcds. *CA: A Cancer J. for Clinicians*, 60(5):301–316, 2010.

[193] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction*, 5(3):207–298, 2013.

[194] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[195] S. Romano, G. Scanniello, M. Risi, and C. Gravino. Clustering and lexical information support for the recovery of design pattern in source code. In *27th IEEE Intl. Conf. on Software Maintenance (ICSM)*, pages 500–503. IEEE, 2011.

[196] T. Ropinski, S. Oeltze, and B. Preim. Survey of glyph-based visualization techniques for spatial multivariate medical data. *Computers & Graphics*, 35(2):392–401, 2011.

[197] H. Saini, S. P. Lal, V. V. Naidu, V. W. Pickering, G. Singh, T. Tsunoda, and A. Sharma. Gene masking-a technique to improve accuracy for cancer classification with high dimensionality in microarray data. *BMC Medical Genomics*, 9(3):261–269, 2016.

[198] S. Sayed, M. Nassef, A. Badr, and I. Farag. A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. *Expert Systems with Applications*, 121:233–243, 2019.

[199] J. D. Schaffer and J. J. Grefenstette. Multi-objective learning via genetic algorithms. In *Procs. of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 593–595. Morgan Kaufmann Publishers Inc., 1985.

[200] G. Schuler, M. Boguski, E. Stewart, L. Stein, G. Gyapay, K. Rice, R. White, P. c. Rodriguez-Tomé, A. Aggarwal, E. Bajorek, et al. A gene map of the human genome. *Science*, 274(5287):540–546, 1996.

[201] G. D. Schuler. Sequence mapping by electronic pcr. *Genome research*, 7(5):541–550, 1997.

[202] G. D. Schuler. Electronic pcr: bridging the gap between genome mapping and genome sequencing. *Trends in biotechnology*, 16(11):456–459, 1998.

[203] D. Selivanov and Q. Wang. text2vec: Modern text mining framework for r. *Computer software manual](R package version 0.4. 0). Retrieved from https://CRAN. R-project. org/package= text2vec*, 2016.

[204] S. Sengupta, A. Singh, H. A. Leopold, T. Gulati, and V. Lakshminarayanan. Ophthalmic diagnosis using deep learning with fundus images–a critical review. *Artificial Intelligence in Medicine*, 102:101758, 2020.

[205] D. Shah et al. IoT based biometrics implementation on Raspberry Pi. *Procedia Computer Science*, 79:328–336, 2016.

[206] N. Singh, D. Gandhi, and K. P. Singh. Iris recognition system using a Canny edge detection and a circular hough transform. *Intl. Journal of Advances in Engineering & Technology*, 1(2):221, 2011.

[207] A. Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[208] S. N. Sivanandam and S. N. Deepa. *Introduction to genetic algorithms*. Springer Berlin Heidelberg New York, 2008.

[209] M. Sonka, V. Hlavac, and R. Boyle. Image pre-processing. In *Image Processing, Analysis and Machine Vision*, pages 56–111. Springer, 1993.

[210] L. Steels. The artificial life roots of artificial intelligence. *Artificial life*, 1(1_2):75–110, 1993.

[211] E. A. Stewart, K. B. McKusick, A. Aggarwal, E. Bajorek, S. Brady, A. Chu, N. Fang, D. Hadley, M. Harris, S. Hussain, et al. An sts-based radiation hybrid map of the human genome. *Genome research*, 7(5):422–433, 1997.

[212] W. V. Stoecker, K. Gupta, R. J. Stanley, R. H. Moss, and B. Shrestha. Detection of asymmetric blotches (asymmetric structureless areas) in dermoscopy images of malignant melanoma using relative color. *Skin Research and Technology*, 11(3):179–184, 2005.

[213] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Procs. of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[214] N. N. Sultana and N. B. Puhan. Recent deep learning methods for melanoma detection: a review. In *International Conference on Mathematics and Computing*, pages 118–132. Springer, 2018.

[215] S. Sun, Q. Peng, and X. Zhang. Global feature selection from microarray data using lagrange multipliers. *Knowledge-Based Systems*, 110:267–274, 2016.

[216] The RNA degradation pathway. https://www.genome.jp/kegg-bin/show_pathway?ko03018. [Online].

[217] Y. Tong, W. Lu, Y. Yu, and Y. Shen. Application of machine learning in ophthalmic imaging modalities. *Eye and Vision*, 7:1–15, 2020.

[218] A. S. Toor, H. Wechsler, and M. Nappi. Biometric surveillance using visual question answering. *Pattern Recognition Letters*, 126:111–118, 2019.

[219] M. T. B. Toossi, H. R. Pourreza, H. Zare, M.-H. Sigari, P. Layegh, and A. Azimi. An effective hair removal algorithm for dermoscopy images. *Skin Research and Technology*, 19(3):230–235, 2013.

[220] H. Tsao, J. M. Olazagasti, K. M. Cordoro, J. D. Brewer, S. C. Taylor, J. S. Bordeaux, M.-M. Chren, A. J. Sober, C. Tegeler, R. Bhushan, et al. Early detection of melanoma: reviewing the abcdes. *J. of the American Academy of Dermatology*, 72(4):717–723, 2015.

[221] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018.

[222] S. E. Umbaugh, R. H. Moss, and W. V. Stoecker. Applying artificial intelligence to the identification of variegated coloring in skin tumors. *IEEE Engineering in Medicine and Biology Magazine*, 10(4):57–62, 1991.

[223] H. J. Vala and A. Baxi. A review on Otsu image segmentation algorithm. *Intl. Journal of Advanced Research in Computer Eng. & Tech.*, 2(2):387–389, 2013.

[224] I. Van der Ploeg. Normative assumptions in biometrics: On bodily differences and automated classifications. In *Innovating Government*, pages 29–40. Springer, 2011.

[225] I. Van der Ploeg. Security in the danger zone: Normative issues of next generation biometrics. In *Second generation biometrics: The ethical, legal and social context*, pages 287–303. Springer, 2012.

[226] C. N. Vasconcelos and B. N. Vasconcelos. Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation. *arXiv preprint arXiv:1702.07025*, 2017.

[227] E. Vocaturo, E. Zumpano, and P. Veltri. Features for melanoma lesions characterization in computer vision systems. In *Information, Intelligence, Systems and Applications (IISA)*, pages 1–8. IEEE, 2018.

[228] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg. Real-time detection and tracking for augmented reality on mobile phones. *IEEE Trans. on Visualization and Computer Graphics*, 16(3):355–368, 2010.

[229] D. Wagner and D. Schmalstieg. First steps towards handheld augmented reality. In *7th IEEE Intl. Symposium on Wearable Computers*, pages 127–135, 2003.

[230] P. Wang. On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37, 2019.

[231] Y. Wen, L. Chen, L. Qiao, Y. Deng, S. Dai, J. Chen, and C. Zhou. Symptom and pathology report generation for ophthalmic diseases in fundus images. In *IEEE Intl. Conf. on Bioinformatics and Biomedicine (BIBM)*, pages 349–356, 2020.

[232] K. Weskamp and S. J. Barmada. RNA degradation in neurodegenerative disease. *RNA Metabolism in Neurodegenerative Diseases*, pages 103–142, 2018.

[233] V. L. West, D. Borland, and W. E. Hammond. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*, 22(2):330–339, 2015.

[234] C. A. White and L. A. Salamonsen. A guide to issues in microarray analysis: application to endometrial biology. *Reproduction*, 130(1):1–13, 2005.

[235] L. Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980.

[236] D. Wittkower. Technology and discrimination. 2018.

[237] P. Wu and D. Wang. Classification of a DNA microarray for diagnosing cancer using a complex network based method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3):801–808, 2018.

[238] C. Xie, Y.-K. Leung, A. Chen, D.-X. Long, C. Hoyo, and S.-M. Ho. Differential methylation values in differential methylation analysis. *Bioinformatics*, 35(7):1094–1097, 2019.

[239] G. Xie and W. Lu. Image edge detection based on OpenCV. *Intl. Journal of Electronics and Electrical Engineering*, 1(2):104–6, 2013.

[240] Q. Xu, Y. Park, X. Huang, A. Hollenbeck, A. Blair, A. Schatzkin, and H. Chen. Diabetes and risk of parkinsonâs disease. *Diabetes care*, 34(4):910–915, 2011.

[241] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews*, 119(18):10520–10594, 2019.

[242] J. J. Yasmin and M. M. Sadiq. An improved iterative segmentation algorithm using canny edge detector with iterative median filter for skin lesion border detection. *International Journal of Computer Applications*, 975:8887, 2012.

[243] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2016.

[244] S. Yue. Human motion tracking and positioning for augmented reality. *Journal of Real-Time Image Processing*, pages 1–12, 2020.

[245] P. S. Zarrin and C. Wenger. Pattern recognition for copd diagnostics using an artificial neural network and its potential integration on hardware-based neuromorphic platforms. In *International Conference on Artificial Neural Networks*, pages 284–288. Springer, 2019.

[246] Y. Zhang, J. Szustakowski, and M. Schinke. Bioinformatics analysis of microarray data. *Cardiovascular Genomics*, pages 259–284, 2009.

[247] T. Zhao, Y. Liu, G. Huo, and X. Zhu. A deep learning iris recognition method based on capsule network architecture. *IEEE Access*, 7:49691–49701, 2019.

[248] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, 2004.

APPENDIX A: THRESHOLD SETTING

For the choice of a biometric threshold, various publications on the subject were examined, to study the literature and the choices made in similar case studies.

First of all, the study of John Daugman himself [56] was examined, who observes how, out of a total of more than 7000 comparisons on images of the same irises, the maximum value for the observed Hamming distance is equal to 0.327, while from the comparison of more than 9 million different irises, it emerged that the probability of a false match between two independent encodings, using an acceptance threshold of 0.33, is 1 case in 4 million, as shown in the Figure 6.1 and from Figure 6.2.

Another case study examined is that of Libor Masek [153], author of the open-source implementation of the Daugman algorithm. In his publication, two image datasets were examined, the CASIA-a and the LEI-a, as show in the Table 6.1.

Table 6.1: Datasets used by Libor Masek for testing

Set Name	Super Set	Number of Eye Images	Possible Intra-Class Comparison	Possible Inter-Class Comparisons
CASIA-a	CASIA	624	1679	192,699
LEI-a	LEI	75	131	2646

Starting from these intraclass and interclass comparisons, it was possible to arrive at a detailed analysis of False Acceptance and False Rejection Rates, using an ever-increasing threshold, as shown in the following tables 6.2 and 6.3:

For the LEI-a dataset, perfect identification is guaranteed by the 0.4 thresholds, with a FAR and a FRR of 0.0%. In CASIA-a, on the other hand, it was not possible to identify a perfect threshold in the same way, due to the overlapping between the classes, but also in this case, with an acceptance threshold of 0.4 an excellent level of reliability is obtained, with a FAR and FRR of 0.005% and 0.238% respectively, very close to perfection. Libor Masek, therefore, identifies 0.4 as the biometric accep-

HD criterion	Odds of false match
0.26	1 in 10^{13}
0.27	1 in 10^{12}
0.28	1 in 10^{11}
0.29	1 in 13 billion
0.30	1 in 1.5 billion
0.31	1 in 185 million
0.32	1 in 26 million
0.33	1 in 4 million
0.34	1 in 690,000
0.35	1 in 133,000
0.36	1 in 28,000
0.37	1 in 6750
0.38	1 in 1780
0.39	1 in 520
0.40	1 in 170

Figure 6.1: Probability of false matches with increasing HD threshold.

tance threshold. Most of the remaining available case studies were often based on the previous two, in particular the implementation of Masek, through the use of the CASIA-a dataset for comparison in most cases, consequently obtaining very similar results, therefore the studies of John Daugman and Libor Masek were considered for the choice of a biometric threshold. For consistency in the choice, the result taken into consideration was the minimum of the two considered, and therefore that of John Daugman, with an acceptance threshold of 0.33, which showed excellent results in terms of reliability, even for images captured in conditions not optimal.

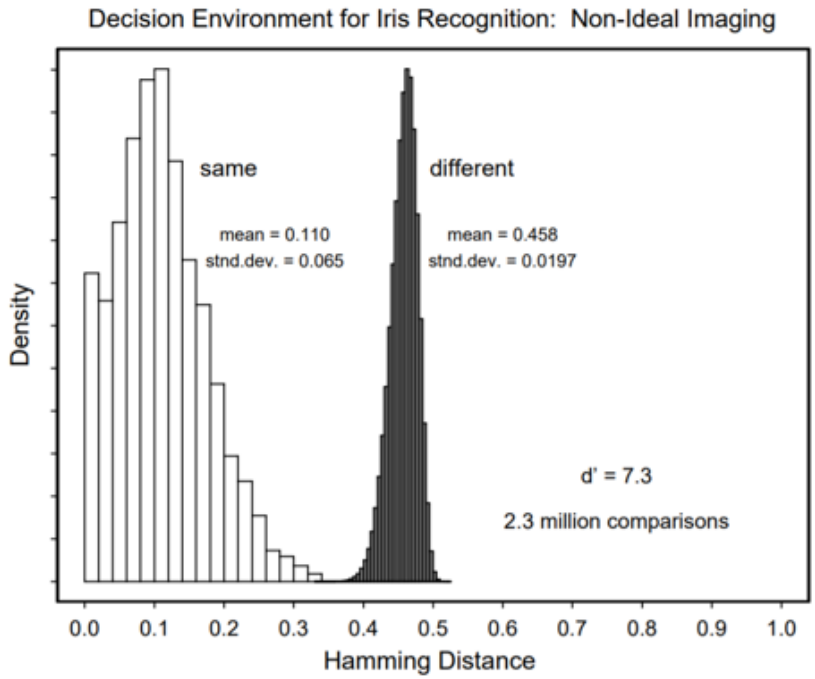


Figure 6.2: Results of intra-class and inter-class matching on non-ideal images.

Table 6.2: False Acceptance and False Rejection Rate for dataset CASIA-a

Threshold	FAR%	FRR%
0.20	0.000	99.047
0.25	0.000	82.787
0.30	0.000	37.880
0.35	0.000	5.181
0.40	0.000	0.238
0.45	7.599	0.000
0.50	99.499	0.000

Table 6.3: False Acceptance e Rejection Rate for dataset LEI-a

Threshold	FAR%	FRR%
0.20	0.000	74.046
0.25	0.000	45.802
0.30	0.000	25.191
0.35	0.000	4.580
0.40	0.000	0.000
0.45	2.494	0.000
0.50	92.819	0.000