Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

Ciclo 34

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

APPLYING MACHINE LEARNING: A MULTI-ROLE PERSPECTIVE

Presentata da: Alessia Angeli

Coordinatore Dottorato
Prof. Daniele Bonacorsi

Supervisore
Prof. Gustavo Marfia

Co-supervisore
Prof. Marco Roccetti

Esame finale anno 2023

Declaration

I declare that, except where stated otherwise by specific reference to the work of others, this dissertation has been composed solely by myself and it has not been submitted, in whole or in part, in any previous application for any other degree or any other university. The work presented is entirely my own except where specified, in the text and Acknowledgements, regarding research collaborations with others. The collaborative contributions have been indicated clearly and due references have been provided on all supporting literature and resources.

Parts of this work have been published in [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13], or submitted in [14].

January 2023

Abstract

Machine (and deep) learning technologies are more and more present in several fields. It is undeniable that many aspects of our society are empowered by such technologies: web searches, content filtering on social networks, recommendations on e-commerce websites, mobile applications, etc., in addition to academic research. Moreover, mobile devices and internet sites, e.g., social networks, support the collection and sharing of information in real time. The pervasive deployment of the aforementioned technological instruments, both hardware and software, has led to the production of huge amounts of data. Such data has become more and more unmanageable, posing challenges to conventional computing platforms, and paving the way to the development and widespread use of the machine and deep learning. Nevertheless, machine learning is not only a technology. Given a task, machine learning is a way of proceeding (a way of thinking), and as such can be approached from different perspectives (points of view). This, in particular, will be the focus of this research. The entire work concentrates on machine learning, starting from different sources of data, e.g., signals and images, applied to different domains, e.g., Sport Science and Social History, and analyzed from different perspectives: from a non-data scientist point of view through tools and platforms; setting a problem stage from scratch; implementing an effective application for classification tasks; improving user interface experience through Data Visualization and eXtended Reality. In essence, not only in a quantitative task, not only in a scientific environment, and not only from a data-scientist perspective, machine (and deep) learning can do the difference.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

"A computer program is said *to learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Tom Mitchell [17]

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience [17, 18]. A class of machine learning computational models that have gained particular interest in the past decade, deep learning, comprises different processing layers, which learn data representations with multiple levels of abstraction, and an increasing number of applications employ these techniques [17, 19]. Machine and deep learning technologies power many aspects of modern society, e.g., web searches, content filtering on social networks, and recommendations on e-commerce websites, and are more and more present in consumer products, e.g., cameras and smartphones. Such technologies can identify objects in images, transcribe speech into text, match new items, posts, or products with users' interests, select relevant search results, etc., to bring some examples [19].

There is no single perspective for research in the field of machine learning. To better analyze this concept, it is possible to start from the elements present in its definition: the task T, the experience E, and the performance measure P;

without forgetting those who interface with these algorithms, developers, domain experts, and users. These elements are all necessary within a machine-learning pipeline. Correctly setting a task, starting from a "good" dataset, and selecting and developing a performing algorithmic strategy are all fundamental steps in a machine learning approach. The work carried out for this Thesis adopts a holistic approach as it involves, as far as possible, all the steps required in the adoption of machine (and deep) learning in its entirety, from its different perspectives, but always in connection with real-world and pragmatic applications.

Firstly, we considered the point of view of problem domain experts that are non-data scientists. In fact, machine learning algorithms solve domain-specific problems in a variety of contexts but their full benefit may be achieved when all categories of interested parties, i.e., including domain experts, who not necessarily be data scientists, will be empowered with their use. Following this, some recent efforts and developments in the area of data science are moving in the direction of making data science itself accessible to non-expert practitioners [20, 21, 22].

Secondly, we focused on a task T defined from scratch, considering all its aspects. The individuation of a problem, the setting of the experiment, the recruitment of the participants, the data collection, and the statistical analysis. For the specific task we addressed, we experienced the necessity of adopting classical analysis methodologies while waiting to collect enough data to apply machine learning algorithms. Data is an integral part of any machine learning solution [23].

Third, we applied machine learning algorithms and models to specific applications and, therefore, on the experience E and the performance measure P. The algorithms (and their architecture) are the essence of machine learning. Nowadays, different learning algorithms, for different applications, in different contexts, exist [24, 25, 26, 27, 28].

Then, we concluded by considering the point of view of users, those who effectively use the developed applications. In particular, how machine learning can improve the interface user experience through eXtended Reality. Extended Reality, e.g., Augmented Reality and Mobile Augmented Reality, is transforming into a technology that may be available in a variety of contexts, expanding from an only

academic or highly specialized technology to an everyday one [29, 30, 31, 32, 33].

The following is a detailed overview of the research presented in this Thesis, in addition to its structure. Importantly, in each Chapter, an additional and specific Introduction, along with Related Works and Contributions is reported.

Chapter 2 In this Chapter, machine learning is focused on and analyzed from a non-data scientist point of view. The focus is to verify how it is possible, without any specific customization, to employ machine learning algorithms provided by different publicly accessible tools and platforms (namely, Weka, Orange, Ludwig, and Knime), to solve a classification problem obtaining significant results and performance that would have been deemed remarkable until not long ago. Finally, we discuss the possible issues and opportunities posed by such an approach, including an additional comparison considering a programming language (namely Python). To do this, we set in the Sports Science area, considering a Human Activity Recognition classification task: to recognize if a unilateral dumbbell bicep curl (a weightlifting exercise) is performed correctly or not, considering the four most common mistakes (for a total of five possible classes).

Chapter 3 In this Chapter, the stage for a machine learning task was set from scratch. In particular, the stage set in the present work wants to explore the distinctive contribution of motor planning and control to Human Reaching Movements in the Psychological field, considering the selection/inhibition of a prepotent response. To do this a portable and low-cost 3-axis wrist-worn accelerometer was utilized to collect raw acceleration data as a starting point for successive kinematics analysis regarding the Reaction Time, Movement Duration, and Time to Peak Velocity of the examined movements.

Chapter 4 In this Chapter, the research started from a collection of vernacular images (the IMAGO collection), belonging to Family Photo Albums, and focused on the effective application of machine learning models, i.e., Convolutional Neural Network and Vision Transformed-based ones, considering different classification tasks, i.e., the dating and the socio-historical context classification, of an image.

This is to present the design and implementation of a multimedia application that, resorting to deep learning models, could assist socio-historians in their cataloging work. Then, always considering these Socio-Historical aspects in the Cultural Heritage environment, we study the relations between quantitative methods and qualitative analyses, including both their potentials and limits. Again, exploiting cross-dataset experiments, we aim to show how deep learning models could reveal their resources, not only in terms of their performance but also in terms of their possible applications to intercultural research.

Chapter 5 In this Chapter, the usage of machine learning to improve user interface experience was explored, through Data Visualization, Virtual and (Mobile) Augmented Reality. Three different (Mobile) Augmented Reality applications and systems in different contexts are proposed and presented in detail. From Cultural Heritage through Family Photo Albums with a Collaborative Photo Environment system, to the Wine Domain with an Augmented Wine Recognition system, passing from an Artisan Work environment like the locksmith one with a Mobile Key Recognition application proposal. In these applications/systems, different components were implemented and/or exploited, among these Optical Character Recognition modules, database structures, search algorithms, machine learning models, and client-server paradigms, in addition to eXtended Reality guided interfaces.

Chapter 6 In this Chapter, a discussion regarding the entire work is reported, with a view to possible future works.

Chapter 2

Machine learning from a non-data scientist point of view

Considering a Human Activity Recognition task in the Sport Science area

2.1 Introduction

Data science has become more and more powerful as the development of algorithms and computing power has made huge progress. In addition, machine learning is nowadays integrated into many areas of everyday life, and this often happens without the end users understanding or even noticing it [18, 19, 24, 25, 34, 35]. However, to harness their full potential, the great majority of tools that are today available to this end require expert data scientists to guarantee the application of the most appropriate algorithms to the context of use [36, 37]. In fact, researchers have used (and continue to use) machine learning algorithms to solve domain-specific problems in a variety of contexts. Nevertheless, the maximum benefit of applying such technologies for the social good may be achieved only when, also non-specialist data scientists will be empowered with their use, capable of understanding and interacting with data, and such a goal may be obtained as intuitive data-human tools and interfaces will be widely available. Some recent efforts and developments in the area of data science are moving in the direction of making

data science itself accessible to non-expert practitioners [20, 38, 21, 22]. For several years researchers have tried to lower the barrier of machine learning through the creation of new systems [39, 40, 41, 42], and now times have perhaps become mature.

This project takes a step along this path. In fact, in this work, we do not want to address data science only in the "classical" way. Focusing on such perspective, we started exploring the data-human connection that may be possible to establish utilizing available tools, with little or no background knowledge of how a data science pipeline works. To this aim, we picked an exemplar scenario, and off-the-shelf machine learning algorithms are put to good use to obtain meaningful results in the field of qualitative activity recognition. To know that qualitative activity recognition studies aim at determining the quality of execution of a given movement, rather than recognizing the movement itself (which may have been implemented at an earlier stage of the process or simply be known a priori). With this information in hand, it would be possible to give (real-time) feedback on the quality of the movement that has been performed, information that is key in many different domains [43, 44, 45, 46]: *Human Activity Recognition (HAR) has emerged as a key research area in the last years and is gaining increasing attention by the pervasive computing research community* [47]. To pursue such an aim, we started from an existing qualitative activity recognition dataset, containing raw data and computed parameters obtained from Inertial Measurement Units (IMUs), and in the past analyzed utilizing feature selection techniques and custom machine learning paradigms [15]. The IMUs were attached to athletes while performing a weightlifting exercise with dumbbells. The focus of the researchers that created and first studied such a dataset was to devise a machine learning model capable of identifying the correct way of lifting dumbbells, as well as four typical mistakes. To this end, they exploited a feature selection approach and a custom model for the classification of weightlifting movements.

In this project, however, we verify how it is possible, without any specific customization, i.e., without changing the default settings and/or performing any type of feature selection, to employ machine learning algorithms, provided by differ-

ent publicly accessible platforms (namely, Weka, Orange, Ludwig, and Knime), to solve a classification problem obtaining significant results and performance that would have been deemed remarkable until not long ago. Nevertheless, since not all of the utilized platforms and algorithms have led to satisfactory results, we finally discuss the possible issues and opportunities posed by such an approach, including an additional comparison considering a programming language (Python). More in detail, we considered a few of the most prominent simplified data-human interface machine learning platforms, devised to be used by non-expert users:

- **Weka**, "Waikato Environment for Knowledge Analysis", developed at the University of Waikato in New Zealand since 1993 [48];
- **Orange**, developed at the University of Ljubljana since 1996 [49];
- **Ludwig**, a deep learning software released in early 2019 by Uber [50];
- **Knime**, initially developed at the University of Konstanz, which specializes in pharmaceutical applications [51].

In addition, this work refers to the respective publications [1, 2], including a taxonomy of the considered learning platforms to aid the non-expert user with a map of the strengths and weaknesses of such tools.

The remainder of this Chapter is organized as follows. In Section 2.2 the most relevant literature for this work, including data-human interfaces (Section 2.2.1) and activity recognition (Section 2.2.2), is reviewed. In Section 2.3 the main contributions are highlighted. Section 2.5 introduces the platforms and the programming language that have been here employed as user-data interfaces presents the models considered for the experiments, and shows the results obtained with the different approaches. Section 2.6 reports a discussion to summarize the findings and draw conclusions.

2.2 Related works

The importance of granting the opportunity of profitably utilizing machine learning algorithms to all is gaining thrust, as witnessed by the increasing number of works that are appearing in the literature on this theme. In this Section, we review a sample of the most relevant scientific works concerned with the importance of easy-to-use data-human interfaces, in Section 2.2.1, as well as a few of those related to the domain of Human Activity Recognition (HAR), in Section 2.2.2.

2.2.1 Data-human interfaces

The ever-increasing amount of collected data, in several different domains, is leading many non-experts in data science to face the problem of diagnosing and solving problems with the use of machine learning [26, 52, 53, 54, 55, 56, 27, 57]. To support such needs, the research community has studied and developed platforms that aim at easing the use of the existing models and algorithms by non-experts. In the following, we present three different contributions and alternatives to those exploited in this project, where the problem of empowering non-data scientists with adequate data analysis tools has been considered.

In particular, Patel et al. [58] firstly discussed the difficulties of applying machine learning algorithms for non-data scientists, pointing out the need for tools that could let a wider community of developers effectively use them. Secondly, they worked on tools that could lower the applicability barrier of such algorithms. To pursue a such objective, the author created Gestalt, a prototype integrated development environment. Gestalt provides explicit support for connecting the principal steps in a pipeline of a machine learning algorithm and an interactive graphical interface through which developers can quickly sort and filter examples to drill down into the data they need.

Again, analyzing situations in which non-expert practitioners may use machine learning algorithms, Yang et al. [59] investigated how a such set of users build machine learning solutions along with the problem they encounter. The authors concluded that, even though it is challenging, a machine learning tool for non-experts

should be both easy to use and robust. To advance on this insight, the authors discussed design implications and created a sensitizing concept to demonstrate how designers might guide non-experts to easily build robust solutions.

Finally, Chen et al. [60] focused on studying how visualization could improve learning machines in order to be more accessible to more users, as it happens in many complex data-oriented domains. In detail, they employed a visual interface to engage practitioners in a design exercise that explored how they would carry out multi-step diagnosis.

Several studies have tried, hence, to simplify the language of data science by linking it to visual metaphors to facilitate the interaction model between humans and data. Unlike the works here discussed, the present contribution considers different visual and non-visual programming approaches, aiming at showing where critical issues may emerge in the process of approaching data science with off-the-shelf solutions. Considering a real-world scenario, we exhibit the potentials and limits that may arise when employing a specific set of machine learning algorithms with the individuated platforms.

2.2.2 Human activity recognition

HAR is an emerging field of research in pervasive computing and human-computer interaction, due to the enormous improvement of sensor technologies and the constantly increasing computing power. It aims at recognizing human activity based on the data obtained from different sensor sources, such as video cameras, wearable sensors (e.g., accelerometer, magnetometer, gyroscope) or ambient sensors (e.g., radar, sound sensors, pressure sensors, temperature sensors) [61]. Many works have so far explored this approach, utilizing both supervised and unsupervised learning methodologies [15, 1, 27, 62, 63, 64, 65, 66, 67, 68]. HAR has already become part of our everyday life when we think, for instance, of smartphones or smartwatches that are capable of detecting basic activity states with the help of a simple built-in accelerometer [62, 69, 70]. The application areas of HAR are manifold, including the recognition of daily life activities [71, 72, 73], the assessment of skill and performance in sports [15, 74, 75, 76], the monitor-

ing of long-term health conditions for disease diagnostics [77, 78, 26, 52, 79], and training of personnel in industrial and maintenance processes [80, 81, 44]. Several researches have focused on a qualitative assessment of human activities. The works in this area are more concerned with *how* an activity is performed, rather than with *which* that activity was performed.

In [76], for instance, Ladha et al. developed the skill assessment platform ClimbAX for climbing using tri-axial accelerometers. Analyzing the acceleration patterns of competitive climbers, they were able to find a correlation between the sensor-data predicted scores, based on performance attributes derived from the raw acceleration data, and the competition scores.

In [82], Khan et al. proposed a generalized skill assessment framework using a hierarchical and stochastic rule induction method. The framework was tested in the context of surgical skill assessment of medical students.

In [15], instead, Velloso et al. investigated the feasibility of automatically assessing the quality of the execution of weightlifting exercises. In particular, four Inertial Measurement Units (IMUs) were used to track the motion of a unilateral dumbbell biceps curl. The participants were asked to perform biceps curl repetitions in five different ways (5 different classes): one correctly and four incorrectly, according to the most ways common mistakes. Then, the gathered data were used as a training dataset for a machine learning algorithm: a 10-fold Random Forest. In the first step, the authors processed the raw data, computing derived features per each, in correspondence to each curl movement, for a total of 96. In a second step, utilizing a correlation-based feature selection algorithm, they determined the 17 most relevant features necessary to describe each curl repetition.

Concluding, in this work, a different approach has been adopted compared to the works discussed in this Section, building on the raw dataset published in [15]. Taking a non-expert data scientist's perspective, only raw data is used. In fact, unlike the previously discussed approach, no further analysis of the dataset and no feature selection algorithm have been implemented. Such an approach brought us to consider all the available sensor data as provided, i.e., the raw data was

not grouped/analyzed in correspondence with each curl movement and no feature selection algorithm was employed. In other words, the sensor data has been considered as it is, and such a dataset has been simply manipulated and employed without requiring specific data science expertise.

2.3 Contributions

The main contribution of this project amounts: (i) to show how, without the need for any specific training in data analysis, existing and newly released data science platforms may now be put to good use to obtain interesting results; (ii) to notice that such results are comparable to those obtained in past studies [15], where custom-designed algorithms were instead engineered for the same purpose. A situation that is possible also thanks to the increasing availability of computing resources supporting the use of raw datasets as a whole. This is simply done by putting such interfaces to good use, without changing the default model settings and/or performing any type of feature selection. As a final term of comparison, we also exploited the use of the deep learning libraries provided by an interpreted programming language, namely Python [83].

In summary, this research provides: (a) an assessment and comparison of the results obtainable by non-data scientists when resorting to easy-to-use, off-the-shelf, visual and non-visual-based data science platforms applied to a specific and well-investigated problem; (b) the verification of the possible pitfalls a non-data scientist could fall on with the use of the considered platforms (i.e., Weka, Orange, Ludwig, and Knime); (c) a final analysis of the chosen dataset, which serves as a further term of comparison, based on the use of mainstream Python-based learning libraries.

2.4 Dataset

This project is based on the dataset provided by Velloso et al. [15]. As aforementioned, the dataset contains the raw Inertial Measurement Units (IMUs) data

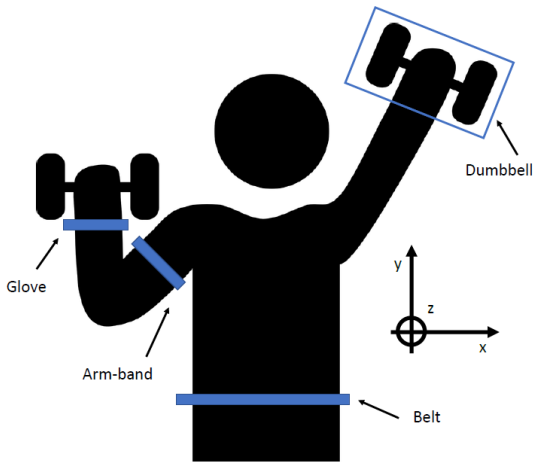


Figure 2.1: IMUs setup by Velloso et al. [15].

of all participants, the extracted features, and the specification related to "how" the bicep curl is performed: the target feature *classe*. In addition, information such as the participants' names is reported, for a total of 158 features (excluding the target one).

More in detail, four IMUs were used to track the motion of a unilateral dumbbell biceps curl (a scheme is reported in Fig. 2.1). The participants were asked to perform one set of 10 repetitions of the biceps curl in five different ways: correctly (*classe* value A) and incorrectly, according to four different but ways common mistakes (*classe* values B, C, D, E, respectively). Then, the authors processed the raw data, sampled at 45 Hz, provided by the 4 sensors, computing 8 features per each, in correspondence to each curl movement: mean, variance, standard deviation, max, min, amplitude, kurtosis and skewness, generating in total 96 derived features. Considering the `csv` formatted file, this consists of a row for the features description, $39,242$ rows for raw data, and 159 columns.

In contrast to [15], who used a sliding window approach [84] to generate their feature set, in the proposed method the user is placed in front of the raw data and no features selection algorithm is used. The only step that is taken, when processing the original dataset, was to ignore all extracted features and information

that did not directly originate from the sensors, such as information about the participants (e.g., participants' names). Important to note that this operation can also be performed without the need for any data science-specific tool (e.g., with a program like Excel). Consequently, only the raw Euler angles (i.e., roll, pitch, and yaw), the raw data provided by the accelerometer, gyroscope, and magnetometer, and the total acceleration were used as inputs. This results in a total of 52 input features:

$$4IMUs \cdot (3EulerAngles + (3Sensors \cdot 3Axis) + Acc_{\text{total}}) = 52.$$

Then, the same procedure was repeated using only the Euler angles and the total acceleration as inputs, resulting in a total of 16 input features:

$$4IMUs \cdot (3EulerAngles + Acc_{\text{total}}) = 16.$$

Such a feature set was identified to contrast it with one of the comparable dimensions, but different variables, employed in [15]. For both settings, the target feature is the feature *classe*. The label A corresponds to the correctly executed movements, while the other labels B, C, D, and E correspond to different classes of incorrectly executed movements.

2.5 Methods and experiments

Following the purpose of this work, in order to portray the results a classifier achieves, before getting to show the learning platforms and methods used, a few of the most common quantities which are used in literature for such aim are here reported. In particular, considering the different rates of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), the following different quantities can be computed.

Precision The proportion of positive identifications that were actually correct;

$$precision = \frac{TP}{TP + FP}$$

Recall The proportion of actual positives that were identified correctly;

$$recall = \frac{TP}{TP + FN}$$

Accuracy The ratio of the number of correct identification to the total number of input samples;

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

F1 score A function of *precision* and *recall*, in particular:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

F1 score might be a better measure to use than *accuracy* in order to seek a balance between *precision* and *recall*, and there is an uneven class distribution.

ROC curve and AUC The Receiver Operating Characteristic curve (*ROC curve*) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate (TPR) and False Positive Rate (FPR), where:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

AUC, instead, measures the entire two-dimensional area underneath the entire *ROC curve* from $(0, 0)$ to $(1, 1)$, and ranges in value from 0 to 1. A possible interpretation of AUC is the probability that the model ranks a random positive sample more highly than a random negative sample.

The aforementioned metrics are effectively used in data science environments, however, require a mathematical background, which exceeds the one simply necessary to distinguish a TP from a TN (as well as FP and FN). Then, we decided to compute the *Confusion Matrix*: characterized by an interesting visual impact, as it is not summarized by a scalar numeric figure but to a bi-dimensional matrix.

Confusion matrix The *confusion matrix*, or error matrix, is a matrix where each row represents the actual values, i.e., the real values, while each column represents the predicted ones. The element on row i and on column j represents the number of cases in which class i has been classified as class j. In this work, in particular, after having computed the *confusion matrix*, the element (i, j) has been divided by the total number of actual elements of the class related to the row i and multiplied by 100, in order to obtain percentage values and results that may be simpler to interpret. Next to the rows and columns of each *confusion matrix*, we will then report the number of actual and predicted elements, respectively, belonging to the different classes. From now on we will refer to this matrix as *confusion matrix - proportion of actual*.

In the following, we discuss the results obtained by exploiting a Random Forest, other ensemble models, and a neural network with Weka, Orange, Knime, and Python. With Ludwig, we employ its default Encoder-Decoder neural network. In addition, for the specific case of neural networks, we also provide a succinct analysis concerning the parameter values adopted within the learning platforms and the Python libraries.

2.5.1 Weka

The Weka platform, "Waikato Environment for Knowledge Analysis", represents a graphical interface to open source machine learning software. In particular, Weka supports several standard data mining tasks such as data preprocessing, clustering, classification, regression, visualization, and feature selection. Containing a plethora of built-in tools for standard machine learning tasks, it is widely used for teaching, research, and industrial applications.

Models Within this contribution, Weka was employed to use models which fall under the neural network and ensemble model classes. In particular, the default Multi Layer Perceptron (MLP), Random Forest, and Bagging models were built to analyze the aforementioned dataset (Fig. 2.2) [85, 57]. Using this platform, only

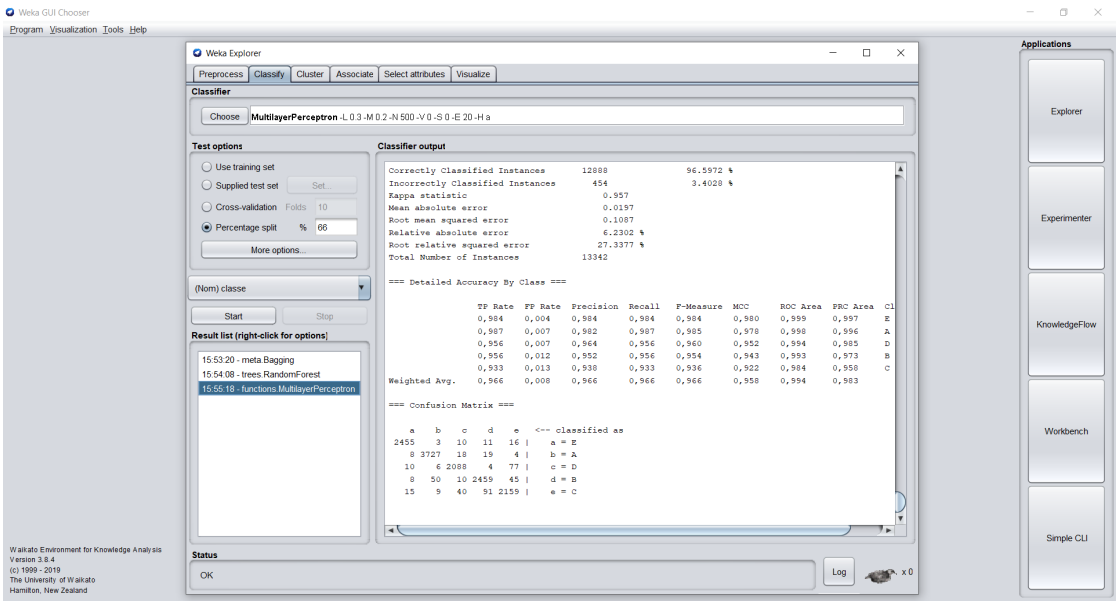


Figure 2.2: Models in Weka.

the configuration related to the size of the training set was changed concerning the default values and, in particular, was set to the 66% of the entire dataset (the default value takes into account a cross-validation option).

Hardware setup and results The training of the models in Weka was performed with a laptop with: (a) RAM: 12GB; (b) Processor: Intel(R) Core(TM) i7-7500U CPU 2.70GHz-2.90GHz; (c) System: Windows 10 Home (64-Bit).

When 52 input features were considered, 457 s were required to train the models, while with 16 input features this step took 97 s. The *confusion matrices - proportion of actual* for MLP, Random Forest, and Bagging in both settings are shown in Table 2.1.

Discussion With Random Forest and Bagging it is possible to observe excellent results both considering 52 and 16 input features. With MLP excellent results can be observed considering 52 input features, whereas a substantial increase in the error rate is observed with 16 input features.

Multi Layer Perceptron - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**98.70%**	0.50%	0.11%	0.48%	0.21%	**3,776**
B	1.94%	**95.61%**	1.75%	0.39%	0.31%	**2,572**
C	0.39%	3.93%	**93.30%**	1.73%	0.65%	**2,314**
D	0.27%	0.18%	3.52%	**95.56%**	0.46%	**2,185**
E	0.12%	0.44%	0.64%	0.40%	**98.40%**	**2,495**
sum	**3,795**	**2,584**	**2,301**	**2,166**	**2,496**	**13,342**

Multi Layer Perceptron - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**89.96%**	2.67%	2.99%	3.97%	0.40%	**3,776**
B	12.33%	**61.39%**	13.41%	7.70%	5.17%	**2,572**
C	2.77%	10.20%	**75.50%**	9.29%	2.25%	**2,314**
D	1.42%	3.43%	9.61%	**84.53%**	1.01%	**2,185**
E	2.40%	5.17%	7.78%	4.85%	**79.80%**	**2,495**
sum	**3,869**	**2,120**	**2,609**	**2,531**	**2,213**	**13,342**

Random Forest - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**99.97%**	0.03%	0.00%	0.00%	0.00%	**3,776**
B	0.12%	**99.88%**	0.00%	0.00%	0.00%	**2,572**
C	0.00%	0.00%	**99.96%**	0.04%	0.00%	**2,314**
D	0.00%	0.00%	0.37%	**99.59%**	0.05%	**2,185**
E	0.00%	0.00%	0.00%	0.04%	**99.96%**	**2,495**
sum	**3,788**	**2,570**	**2,321**	**2,178**	**2,495**	**13,342**

Random Forest - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**99.95%**	0.03%	0.00%	0.00%	0.03%	**3,776**
B	0.08%	**99.57%**	0.31%	0.04%	0.00%	**2,572**
C	0.00%	0.26%	**99.35%**	0.39%	0.00%	**2,314**
D	0.05%	0.00%	0.09%	**99.82%**	0.05%	**2,185**
E	0.00%	0.04%	0.08%	0.00%	**99.88%**	**2,495**
sum	**3,777**	**2,569**	**2,311**	**2,191**	**2,494**	**13,342**

Bagging - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**99.71%**	0.08%	0.16%	0.00%	0.05%	**3,776**
B	0.70%	**98.41%**	0.51%	0.27%	0.12%	**2,572**
C	0.13%	0.13%	**99.22%**	0.52%	0.00%	**2,314**
D	0.09%	0.41%	1.37%	**99.08%**	0.05%	**2,185**
E	0.00%	0.20%	0.24%	0.24%	**99.32%**	**2,495**
sum	**3788**	**2,551**	**2,351**	**2,168**	**2,484**	**13,342**

Bagging - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**99.07%**	0.69%	0.00%	0.21%	0.03%	**3,776**
B	1.09%	**99.97%**	1.40%	0.31%	0.23%	**2,572**
C	0.17%	1.08%	**98.23%**	0.43%	0.09%	**2,314**
D	0.14%	0.37%	0.96%	**98.44%**	0.09%	**2,185**
E	0.16%	0.92%	0.48%	0.24%	**98.20%**	**2,495**
sum	**3,780**	**2,576**	**2,342**	**2,183**	**2461**	**13,342**

Table 2.1: *Confusion matrices - proportion of actual* respect to MLP, Random Forest, and Bagging with Weka, using 66% of the dataset as training data.

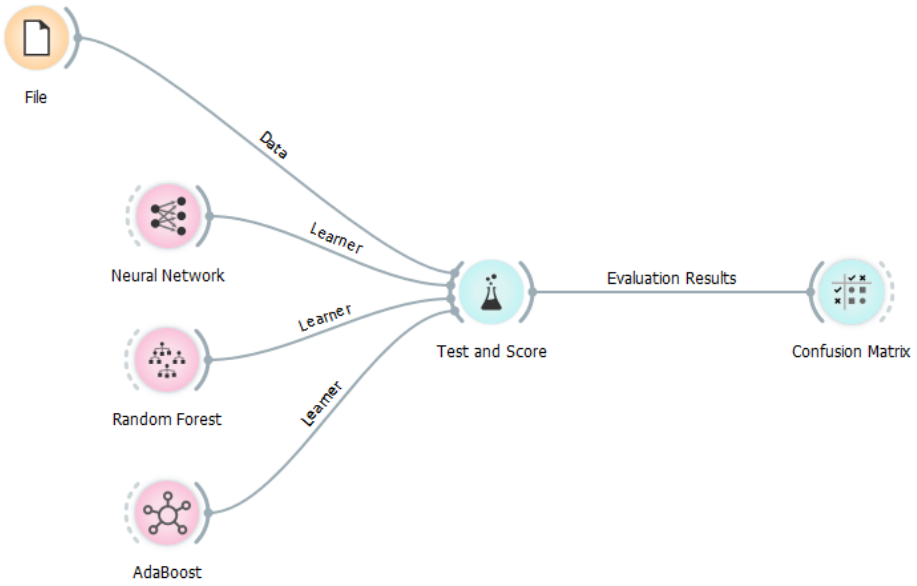


Figure 2.3: Models in Orange.

2.5.2 Orange

Orange is an open-source data visualization, machine learning, and data mining toolkit. In Orange, data analysis is done by stacking components, called widgets, into workflows: there is a large library of widgets and these components communicate with each other in the built models. Combining different widgets enables you to build comprehensive data analysis models as you want, and using the Orange interactive graphical user interface, complex data analytics pipelines can be built focusing on the data analysis part, rather than coding. These characteristics make Orange a useful tool for users with little or no coding experience and knowledge of data science in general.

Models In Orange, the data analysis has been performed resorting to default Neural Network, Random Forest, and AdaBoost models [86, 87]. Fig. 2.3 shows the visual representation of the data analysis pipeline constructed.

Neural Network - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**99.83%**	0.14%	0.01%	0.01%	0.01%	**37,940**
B	0.18%	**99.59%**	0.16%	0.02%	0.06%	**25,820**
C	0.01%	0.25%	**99.27%**	0.44%	0.02%	**23270**
D	0.03%	0.00%	0.51%	**99.22%**	0.23%	**21,870**
E	0.00%	0.05%	0.05%	0.17%	**99.72%**	**24,530**
sum	**37,931**	**25,839**	**23,269**	**21,855**	**24,536**	**133,430**

Neural Network - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**98.04%**	0.96%	0.46%	0.39%	0.15%	**37,940**
B	2.65%	**91.20%**	3.90%	1.03%	1.21%	**25,820**
C	0.58%	3.87%	**90.41%**	4.20%	0.94%	**23,270**
D	0.54%	0.81%	4.78%	**92.63%**	1.23%	**21,870**
E	0.25%	1.03%	1.87%	2.52%	**94.33%**	**24,530**
sum	**38,195**	**25,243**	**23,725**	**22,271**	**23,996**	**133,430**

Random Forest - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**99.94%**	0.04%	0.01%	0.01%	0.01%	**37,940**
B	0.41%	**99.38%**	0.17%	0.01%	0.02%	**25,820**
C	0.01%	0.50%	**99.30%**	0.19%	0.00%	**23,270**
D	0.01%	0.01%	0.81%	**99.09%**	0.08%	**21,870**
E	0.00%	0.04%	0.04%	0.20%	**99.73%**	**24,530**
sum	**38,028**	**25,802**	**23,343**	**21,768**	**24,489**	**133,430**

Random Forest - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**99.78%**	0.12%	0.05%	0.02%	0.03%	**37,940**
B	0.62%	**98.30%**	0.93%	0.09%	0.07%	**25,820**
C	0.06%	0.57%	**98.76%**	0.59%	0.02%	**23,270**
D	0.02%	0.06%	0.50%	**99.32%**	0.09%	**21,870**
E	0.05%	0.13%	0.24%	0.19%	**99.39%**	**24,530**
sum	**38,044**	**25,607**	**23,408**	**21,938**	**24,433**	**133,430**

AdaBoost - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**98.61%**	0.81%	0.20%	0.24%	0.14%	**37,940**
B	1.32%	**96.17%**	1.36%	0.57%	0.59%	**25,820**
C	0.35%	1.26%	**96.71%**	1.24%	0.43%	**23,270**
D	0.43%	0.54%	1.34%	**97.17%**	0.51%	**21,870**
E	0.13%	0.74%	0.62%	0.60%	**97.91%**	**24,530**
sum	**37,965**	**25,729**	**23,378**	**21,923**	**24,435**	**133,430**

AdaBoost - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**98.64%**	0.81%	0.17%	0.18%	0.20%	**37,940**
B	1.18%	**95.91%**	1.53%	0.72%	0.66%	**25,820**
C	0.24%	1.54%	**96.84%**	0.98%	0.40%	**23,270**
D	0.36%	0.77%	0.98%	**97.45%**	0.44%	**21,870**
E	0.18%	0.92%	0.44%	0.61%	**97.85%**	**24,530**
sum	**37,905**	**25,823**	**23,316**	**21,947**	**24,439**	**133,430**

Table 2.2: *Confusion matrices - proportion of actual* respect to Neural Network, Random Forest, and AdaBoost with Orange, using 66% of the dataset as training data and 10 repetitions of train/test.

Hardware setup and results The training of the models in Orange was performed with a laptop with: (a) RAM: 16GB; (b) Processor: Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 1992 Mhz, 4 Core(s), 8 Logical Processor(s); (c) System: Microsoft Windows 10 Home.

The first run with 52 input features took 244 s to train the models using 66% of the dataset, and 10 repetitions of train/test. In the second run with 16 input features, and the same settings, 220 s were required. The *confusion matrices - proportion of actual* for Neural Network, Random Forest, and AdaBoost in both settings are shown in Table 2.2.

Discussion Neural Network and Random Forest both showed optimal reliability when the 52 input features were used. Also, AdaBoost performed very well, thus showing a higher error rate. Interestingly, the *confusion matrix - proportion of actual* of Random Forest and AdaBoost remained almost unchanged as 16 input features were used, while the Neural Network error rate increased by almost ten percentage points. This means that Random Forest proved to be the robust algorithm when handling the considered raw data.

2.5.3 Ludwig

Ludwig is an open-source deep learning toolbox whose models are characterized by a versatile and flexible Encoder-Decoder architecture: it is possible to use the default model, without setting any parameters or creating a personalized model changing these default parameter values. Furthermore, it is possible to train the model by providing just a tabular file, a `csv`-formatted file, containing the data, and a configuration file, a `yaml`-format file, which defines which columns of the tabular file are input and which are target features. The configuration file also defines the type of features to be used in the model. At the moment of this project, Ludwig contained Encoder-Decoder architecture for features with values enable to the following types: text, numerical, binary, category, set, sequence, image, time series, and bag. With the described characteristics, Ludwig can help users who do not have specific knowledge in coding and/or in data sciences in

csv-format file
filedata.csv

```
roll_belt, pitch_belt, yaw_belt, total_accel_belt, ... , classe
3.70, 41.6, -82.8, 3, 132.0, ... , E
3.66, 42.8, -82.5, 2, 129.0, ... , E
3.58, 43.7, -82.3, 1, 125.0, ... , E
3.56, 44.4, -82.1, 1, 120.0, ... , E
3.57, 45.1, -81.9, 1, 115.0, ... , E
3.45, 45.6, -81.9, 1, 110.0, ... , E
3.31, 46.2, -81.9, 3, 104.0, ... , E
2.91, 46.9, -82.2, 4,  98.6, ... , E
2.31, 47.4, -82.6, 2,  93.2, ... , E
2.00, 47.7, -82.8, 3 ,  88.5, ... , E
...
```

yaml-format file
filedata.yaml

```
input_features:
 -
  name: roll_belt
  type: numerical
 -
  name: pitch_belt
  type: numerical
 -
  name: yaw_belt
  type: numerical
 -
  name: total_accel_belt
  type: numerical
 -
  ...
output_features:
 -
  name: classe
  type: category
```

console

```
> ludwig experiment  --data_csv filedata.csv  --model_definition_file filedata.yaml
```

Figure 2.4: Model in Ludwig.

general to apply deep learning algorithms. Nevertheless, differently from Orange, there is no graphical user interface.

Models In Ludwig the default Encoder-Decoder model has been built: no parameter has been set by the user [88]. The original `csv`-formatted data file was therefore considered and, in relation to this, the `yaml`-formatted file, containing the information related to the features, was built. In the task under analysis, considering the feature types available in Ludwig, all the input features are numerical, while the target feature is categorical (Fig. 2.4). The dataset containing $39,242$ rows of raw data was divided as follows in training, validation, and test

Encoder-Decoder - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**86.11%**	2.01%	7.68%	3.22%	0.98%	**2,239**
B	12.83%	**59.72%**	13.56%	2.05%	11.84%	**1,512**
C	6.54%	6.99%	**78.66%**	0.90%	6.91%	**1,331**
D	7.54%	4.20%	23.31%	**50.04%**	14.92%	**1,287**
E	5.39%	12.54%	8.82%	3.71%	**69.54%**	**1,428**
sum	**2,383**	**1,274**	**1,478**	**1,850**	**812**	**7,797**

Encoder-Decoder - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**72.58%**	8.04%	6.07%	9.02%	4.29%	**2239**
B	22.35%	**36.24%**	15.34%	10.05%	16.01%	**1512**
C	27.72%	7.74%	**50.04%**	7.44%	7.06%	**1331**
D	14.30%	13.52%	18.73%	**44.44%**	9.01%	**1287**
E	13.17%	29.06%	19.26%	14.57%	**23.95%**	**1428**
sum	**2,704**	**1,420**	**890**	**1,550**	**1,233**	**7,797**

Table 2.3: *Confusion matrices - proportion of actual* respect to Ludwig Encoder-Decoder, using approximately 70% of the dataset as training data, 10% as validation, and 20% as test.

set: $27,5884$ rows for training, $3,857$ rows for validation, and the $7,797$ remaining rows for the test set. Considering the aim of this work, it is important to specify that this data split was made by default by the model, no indication of how to divide the data into the different sets was provided by the user.

Hardware setup and results The training of the model in Ludwig was performed with a laptop with: (a) RAM: 12GB; (b) Processor: Intel(R) Core(TM) i7-7500U CPU 2.70GHz-2.90GHz; (c) System: Windows 10 Home (64-Bit).

For the first run with 52 input features, it took 172 s to train the model with 70% of the dataset and with 55 epochs. Important to note that the default value for the epochs number is 100, but after 5 epochs since the last validation accuracy improvement an early stopping occurs, as happened in this case, in which the best value of accuracy on the validation set is at epoch 50. For the second run with 16 input features, it took 228 s to train the model with 70% of the dataset and with 100 epochs. In Table 2.3 the *confusion matrices - proportion of actual* obtained from these tests are shown.

Discussion Interesting results have been obtained with Ludwig. Although percentage values are lower than those obtained with Weka and Orange, the results reveal to be promising, when focusing the attention on the class of correct movements (i.e., class A). The results obtained utilizing 52 input features are significant, especially considering that in many applications it is sufficient to correctly distinguish "correct" from "incorrect" movements. The percentage of times a correct movement (i.e., class A) is detected then decreases when using 16, instead of 52 input features, as shown in Table 2.3.

2.5.4 Knime

Knime, in particular the Knime Analytics Platform, is an open-source software for creating data science models. Knime attempts to make understanding data and designing data science workflows and reusable components accessible to everyone users open, and continuously integrating new developments. With this platform, it is possible to create visual data science workflows with an intuitive, drag-and-drop style graphical interface, and without the need for coding. To build a data science workflow over $2,000$ nodes are available and there is the possibility to model each step of the analysis to control the flow of data. Therefore, like Orange, Knime is characterized by an interactive graphical user interface.

Models In Knime we adopted the same models used in Orange or, when this was not possible, models that belonged to the same class (e.g., AdaBoost and Boosting both belong to the ensemble model class). Therefore, the default Multi Layer Perceptron (MLP), Random Forest and Boosting models have been used to analyze the dataset [89]. The only configuration of the Knime nodes that have been changed, concerning the default values, is the size of the training set, which was set to 66%, while the default value would have been 10% of the entire dataset. The visual representation of the final data science workflow constructed with Knime is shown in Fig. 2.5.

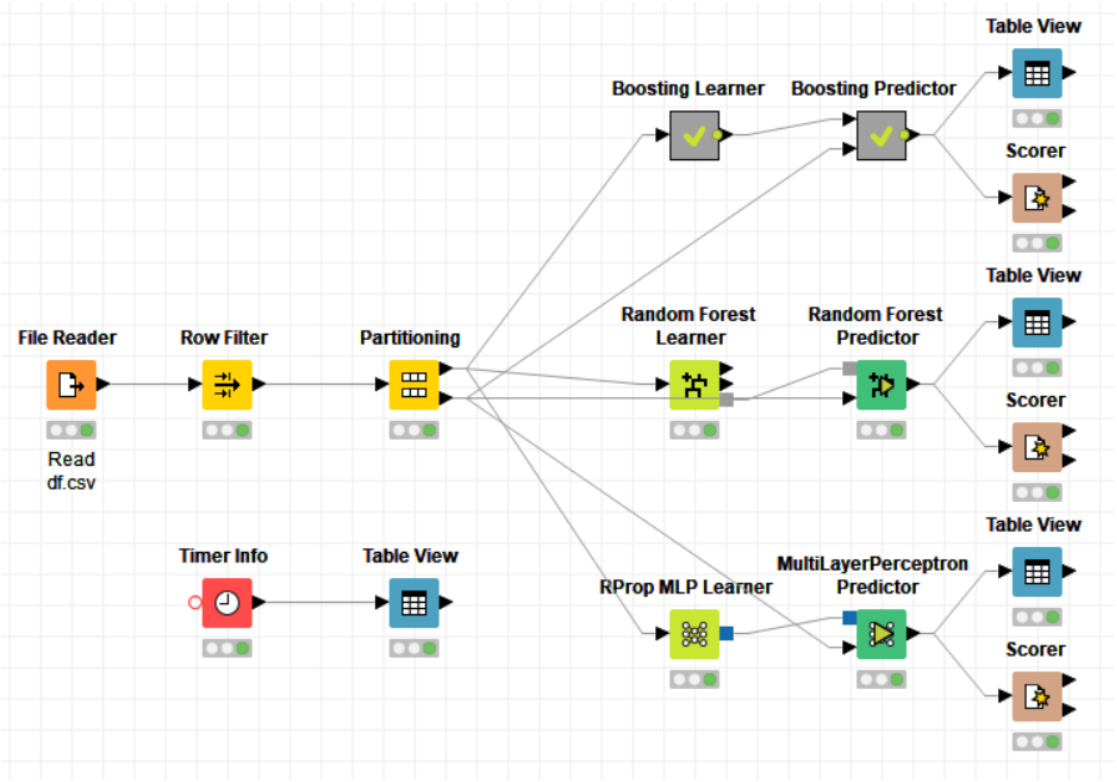


Figure 2.5: Models in Knime.

Hardware setup and results The training of the models in Knime was performed with a laptop with: (a) RAM: 12GB; (b) Processor: Intel(R) Core(TM) i7-7500U CPU 2.70GHz-2.90GHz; (c) System: Windows 10 Home (64-Bit).

When using 52 input features, the training took 83 s, whereas with 16 input features the time required was 32 s. The *confusion matrices - proportion of actual* for MLP, Random Forest, and Boosting in both settings are shown in Table 2.4.

Discussion With the Random Forest algorithm, the obtained results are very similar to those obtained with the Random Forest built in Weka and Orange, for both the 52 and the 16 input features settings. Boosting and MLP were not as successful, instead, in fact, a closer look at the default parameter values reveals that these may not be suitable for the considered classification problem. As shown in Table 2.7 (end of Section 2.5), the main default parameters values for the MLP

Multi Layer Perceptron - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**68.28%**	6.97%	9.24%	10.64%	4.88%	**3,789**
B	11.73%	**37.31%**	13.68%	11.38%	25.89%	**2,565**
C	18.78%	20.51%	**44.05%**	12.03%	4.63%	**2,311**
D	13.41%	8.08%	27.86%	**27.45%**	23.21%	**2,215**
E	8.08%	20.71%	21.77%	15.31%	**34.12%**	**2,462**
sum	**3,818**	**2,384**	**2,872**	**1,958**	**2,310**	**13,342**

Multi Layer Perceptron - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**77.82%**	7.46%	8.44%	3.39%	2.89%	**3,778**
B	18.71%	**45.82%**	15.61%	8.55%	11.31%	**2,608**
C	13.22%	13.61%	**56.35%**	10.82%	6.01%	**2,330**
D	6.38%	16.98%	22.95%	**31.69%**	22.00%	**2,209**
E	4.63%	21.10%	19.32%	9.52%	**45.43%**	**2,417**
sum	**3,989**	**2,679**	**3,013**	**1,533**	**2,128**	**13,342**

Random Forest - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**99.97%**	0.00%	0.00%	0.00%	0.03%	**3,789**
B	0.16%	**99.81%**	0.04%	0.00%	0.00%	**2,565**
C	0.00%	0.22%	**99.61%**	0.17%	0.00%	**2,311**
D	0.00%	0.00%	0.32%	**99.50%**	0.18%	**2,215**
E	0.00%	0.00%	0.00%	0.20%	**99.80%**	**2,462**
sum	**3,792**	**2,565**	**2,310**	**2,213**	**2,462**	**13,342**

Random Forest - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**99.87%**	0.13%	0.00%	0.00%	0.00%	**3,778**
B	0.08%	**99.50%**	0.42%	0.00%	0.00%	**2,608**
C	0.04%	0.17%	**99.44%**	0.34%	0.00%	**2,330**
D	0.00%	0.00%	0.32%	**99.64%**	0.05%	**2,209**
E	0.00%	0.04%	0.08%	0.12%	**99.75%**	**2,417**
sum	**3,776**	**2,605**	**2,337**	**2,212**	**2,412**	**13,342**

Boosting - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**76.56%**	2.96%	13.35%	5.15%	1.98%	**3,789**
B	15.13%	**59.61%**	12.94%	3.94%	8.38%	**2,565**
C	27.65%	10.13%	**50.32%**	8.65%	3.25%	**2,311**
D	21.44%	7.04%	16.03%	**46.00%**	9.48%	**2,215**
E	6.13%	13.12%	6.46%	21.28%	**53.01%**	**2,462**
sum	**4,554**	**2,354**	**2,515**	**2,039**	**1,880**	**13,342**

Boosting - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**66.97%**	9.79%	9.00%	11.04%	3.20%	**3,778**
B	16.10%	**42.06%**	16.53%	13.23%	12.08%	**2,608**
C	12.83%	21.24%	**46.22%**	15.06%	4.64%	**2,330**
D	7.51%	15.84%	8.87%	**55.73%**	12.04%	**2,209**
E	6.91%	25.36%	19.61%	17.19%	**30.33%**	**2,417**
sum	**3,582**	**2,925**	**2,518**	**2,774**	**1,543**	**13,342**

Table 2.4: *Confusion matrices - proportion of actual* respect to MLP, Random Forest, and Boosting with Knime, using 66% of the dataset as training data.

model are the following: 100 maximum number of iterations, 1 hidden layer, and 10 neurons per layer, and in essence may not be sufficient for this task.

2.5.5 Python

As a final experiment, we performed the analysis of the considered data employing Python, an interpreted, high-level, general-purpose programming language that is widely used by the machine learning community [90]. Python, in fact, provides a plethora of libraries available for data science applications. Among the most widely used for modeling, it is possible to find scikit-learn and Keras [91, 92]. Scikit-learn represents a general-purpose library, as it includes both machine and deep learning models. Keras, instead, is more centered on the use of deep learning, hence, neural network-based algorithms. In the following, we explain the differences between using these two libraries from a non-data scientist perspective.

Models In Python, we employed scikit-learn functions to build a Multi Layer Perceptron (MLP), AdaBoost and Random Forest models, resorting as much as possible to the default parameter values. Regarding this, for all the considered models, only one default parameter value was changed: instead of the default value of the 75% of data, we used the 66% of the entire dataset for training, to resemble as much as possible the operational settings before.

Hardware setup and results The training of the models in Python was performed with a laptop with: (a) RAM: 12GB; (b) Processor: Intel(R) Core(TM) i7-7500U CPU 2.70GHz-2.90GHz; (c) System: Windows 10 Home (64-Bit).

The training required 24 s and 16 s when considering the 52 and 16 input features, respectively. The *confusion matrices - proportion of actual* for MLP, Random Forest, and AdaBoost for both settings are shown in Table 2.5.

Discussion In Python, considering the scikit-learn functions, Random Forest and MLP exhibit very good performance, reaching an accuracy close to 100%,

Multi Layer Perceptron - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**99.79%**	0.21%	0.00%	0.00%	0.00%	**3,829**
B	2.87%	**94.10%**	2.37%	0.16%	0.50%	**2,578**
C	0.04%	1.84%	**95.29%**	2.74%	0.09%	**2,336**
D	0.46%	0.28%	6.65%	**90.51%**	2.11%	**2,181**
E	0.00%	0.66%	1.78%	0.70%	**96.86%**	**2,418**
sum	**3,906**	**2,499**	**2,475**	**2,059**	**2,403**	**13,342**

Multi Layer Perceptron - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**97.41%**	1.66%	0.21%	0.37%	0.35%	**3,738**
B	0.93%	**92.05%**	5.60%	0.27%	1.16%	**2,590**
C	0.09%	2.52%	**95.00%**	1.37%	1.03%	**2,341**
D	0.28%	0.41%	7.13%	**91.07%**	1.10%	**2,173**
E	0.20%	0.96%	2.72%	1.92%	**94.20%**	**2,500**
sum	**3,678**	**2,538**	**2,600**	**2,080**	**2,446**	**13,342**

Random Forest - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**100.00%**	0.00%	0.00%	0.00%	0.00%	**3,829**
B	0.08%	**99.88%**	0.04%	0.00%	0.00%	**2,578**
C	0.00%	0.00%	**100.00%**	0.00%	0.00%	**2,336**
D	0.00%	0.00%	0.55%	**99.40%**	0.05%	**2,181**
E	0.00%	0.00%	0.00%	0.04%	**99.96%**	**2,418**
sum	**3,831**	**2,575**	**2,349**	**2,169**	**2,418**	**13,342**

Random Forest - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**99.97%**	0.03%	0.00%	0.00%	0.00%	**3,738**
B	0.23%	**98.80%**	0.89%	0.04%	0.04%	**2,590**
C	0.00%	0.04%	**99.66%**	0.30%	0.00%	**2,341**
D	0.00%	0.00%	0.23%	**99.72%**	0.05%	**2,173**
E	0.00%	0.00%	0.12%	0.16%	**99.72%**	**2,500**
sum	**3,743**	**2,561**	**2,364**	**2,179**	**2,495**	**13,342**

AdaBoost - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**68.06%**	18.41%	8.64%	3.29%	1.59%	**3,829**
B	13.15%	**63.42%**	9.27%	6.90%	7.25%	**2,578**
C	2.05%	6.98%	**80.99%**	7.23%	2.74%	**2,336**
D	3.44%	4.03%	10.13%	**73.59%**	8.80%	**2,181**
E	3.18%	7.82%	4.92%	4.96%	**79.11%**	**2,418**
sum	**3,145**	**2,780**	**2,802**	**2,198**	**2,417**	**13,342**

AdaBoost - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**67.42%**	17.26%	8.56%	4.47%	2.30%	**3,738**
B	12.66%	**62.01%**	11.20%	5.52%	8.61%	**2,590**
C	1.07%	13.46%	**74.93%**	8.80%	1.75%	**2,341**
D	7.18%	3.68%	13.21%	**58.77%**	17.17%	**3,173**
E	5.44%	10.68%	8.04%	6.40%	**69.44%**	**2,500**
sum	**3,165**	**2,913**	**2,852**	**1,953**	**2,459**	**13,342**

Table 2.5: *Confusion matrices - proportion of actual* of MLP, Random Forest, and AdaBoost implemented with Python and scikit-learn functions, using 66% of the dataset as training data.

considering both the 52 and 16 input features, while high error rates are observed when employing AdaBoost.

A layer-by-layer model Resorting then to Keras, we implemented a layer-by-layer neural network and analyzed its behavior when using the default values for the model parameters. Now, it is important to note that it is necessary to set at least a few parameters in Keras: not all of the required parameters have a default value. In Table 2.7 (end of Section 2.5) we specify which parameters are necessary for the MLP model while lacking pre-set default values (i.e., hidden layers number, neurons number in a layer, loss function, and optimizer). In addition, it is possible to observe that the default settings utilized for the number of epochs and the activation function are ill-suited to approach a real-world problem, as they amount to a single epoch and a linear activation function. Hence, we proceeded to adopt the default settings, wherever there were available, deciding to choose as values for the remaining parameters those suggested in [93], a well-known beginners' data science blog. Then, we used the following parameter values: 1 epoch (default value); 2 hidden layers with 32 neurons for each layer; "linear" activation function (default value); "categorical_crossentropy" loss function; "adam" optimizer. The *confusion matrices - proportion of actual* obtained with this setting, considering 52 and 16 input features, respectively, are shown in Table 2.6.

The performance of the described MLP model appears unacceptable. This could be due to the default parameter values: these may not be suitable for the considered classification problem. In particular, we note that considering a linear activation function the resulting MLP behaves as a linear model. Therefore, when using Keras in Python and in particular a model which requires the tuning of different parameters, some knowledge related to how the specific model works are necessary to obtain any significant result.

Finally, we decided to verify the complexity of building in Keras an MLP model which is capable of obtaining significant results. We hence used the same lines of code, simply changing the default values as follows: 40 epochs and 80 epochs considering, respectively, 52 and 16 input features; 2 hidden layers with 32

MLP Keras default - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**37.85%**	56.37%	0.74%	3.38%	1.66%	**3,786**
B	22.66%	**63.96%**	2.63%	4.72%	6.03%	**2,586**
C	32.30%	63.27%	**2.30%**	0.54%	1.59%	**2,393**
D	21.45%	65.62%	0.19%	**9.09%**	3.65%	**2,135**
E	29.85%	58.48%	1.60%	5.24%	**4.83%**	**2,442**
sum	**3,979**	**8,131**	**194**	**585**	**453**	**13,342**

MLP Keras default - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**52.72%**	9.37%	18.93%	5.02%	13.96%	**3,883**
B	50.98%	**16.76%**	16.72%	3.82%	11.70%	**2,589**
C	65.00%	1.36%	**16.97%**	7.08%	9.59%	**2,274**
D	41.04%	16.96%	22.15%	**7.03%**	12.82%	**2,176**
E	54.63%	13.64%	22.23%	4.92%	**4.59%**	**2,420**
sum	**7,060**	**1,528**	**2,574**	**727**	**1,453**	**13,342**

MLP Keras modified - 52 features

actual \ predicted	A	B	C	D	E	sum
A	**98.71%**	0.92%	0.13%	0.21%	0.03%	**3,786**
B	4.80%	**92.81%**	1.89%	0.00%	0.50%	**2,586**
C	0.21%	1.46%	**93.73%**	4.22%	0.38%	**2,393**
D	0.84%	0.28%	4.92%	**93.68%**	0.28%	**2,135**
E	0.45%	0.94%	0.78%	1.80%	**96.03%**	**2,442**
sum	**3,895**	**2,499**	**2,421**	**2,153**	**2,374**	**1,334**

MLP Keras modified - 16 features

actual \ predicted	A	B	C	D	E	sum
A	**94.93%**	2.45%	0.82%	1.44%	0.36%	**3,883**
B	1.43%	**92.08%**	3.24%	1.12%	2.12%	**2,589**
C	0.00%	4.88%	**90.46%**	3.91%	0.75%	**2,274**
D	0.14%	0.46%	4.78%	**93.75%**	0.87%	**2,176**
E	0.04%	1.03%	1.12%	2.44%	**95.37%**	**2,420**
sum	**3,727**	**2,625**	**2,304**	**2,273**	**2,413**	**13,342**

Table 2.6: *Confusion matrices - proportion of actual* respect to MLP implemented layer by layer with Python and Keras, using 66% of the dataset as training data.

neurons for each layer; "relu" activation function and "softmax" activation function for the output layer; "categorical_crossentropy" loss function; "adam" optimizer. The *confusion matrices - proportion of actual* for the MLP models in both settings are shown in Table 2.6. It is possible to observe that, with only a few modifications, it was possible to obtain interesting results: all diagonal values fall above the 90% threshold.

Neural Network/MLP default values

	Weka	Orange	Knime	Python - Scikit-learn	Python - Keras
Epochs	500	200	100	200	1
Hidden layers	1	1	1	1	*no default value*
Neurons per layer	$(attributes + classes)/2$	100	10	100	*no default value*
Activation function	sigmoid	ReLu	clipped logistic	ReLu	linear
Loss function	/	optimized log	/	optimized log	*no default value*
Optimizer	/	adam	rprop	adam	*no default value*

Table 2.7: Default values for principal Neural Network/MLP model parameters in the different considered learning platforms and programming language.

2.6 Discussion and conclusions

In this project, we considered different learning platforms a non-data scientist may invoke to make sense of a collection of raw data. In particular, we treated a classification problem in the area of qualitative activity recognition (i.e., unilateral dumbbell biceps curl in our scenario), and we adopted four different data-human interface learning platforms (Weka, Orange, Ludwig, and Knime) which do not require any coding skills from their users (but, obviously, the raw data), and a programming language as Python. In addition, to follow the purpose of this contribution, simple models, in which the user does not have to set any unnecessary parameter values, have been built.

Main results of our experimental campaign, from the point of view of the choice of the employed default models (a) The Random Forest model has proven to be reliable throughout all the adopted approaches; (b) The other ensemble models (i.e., Bagging, AdaBoost, and Boosting), and neural networks models led to both good and bad results, depending on the platform.

Main results of our experimental campaign, from the point of view of the choice of the employed platforms (a) Weka shows excellent results

(above the 98% threshold) with the Random Forest and Bagging models and both possible feature sets. The MLP model, however, obtained excellent results only with the 52 input features. This is for all the considered classes of movements; (b) Orange obtained very interesting results (above the 90% threshold) with all models and feature sets, for all the considered classes of movements; (c) Ludwig obtained acceptable results (above the 85% threshold) when considering the class of correct movements and the 52 input feature; (d) Knime shows very interesting results (above the 90% threshold) with both feature sets, for all the considered classes of movements, with the Random Forest model. The other models employed in this platform (i.e., MLP and Boosting) led to unacceptable results (below the 80% threshold); (e) The model offered by the scikit-learn library in Python led to controversial results. The MLP and Random Forest models exhibited diagonal values on the *confusion matrices - proportion of actual* which all exceeded the 90% threshold. The AdaBoost model, instead, fell as low as the 58.77% correctly classified for one of the considered cases; (f) The MLP model offered by the Keras library in Python led to very poor and unacceptable results when utilizing the default parameter values. Nevertheless, just changing some of these default values, but not the code structure, we showed how the model could exceed the 90% threshold, for all the classes and both the considered feature sets.

In essence, this work indicates that the safest platform for a non-expert user of data science techniques may be Orange, whereas the best algorithmic choice, among the tested ones, may be the Random Forest. Clearly, further investigations are required, considering additional and diverse datasets as well as further machine learning algorithms. In addition, an observation about the phenomenon of overfitting is necessary regarding the neural network/MLP models, implemented in the considered learning platforms and programming language, to which an in-depth analysis concerning the different default settings was carried out. In particular, if a model has a default setting characterized by parameter values that lead it to have high complexity, it is possible that the model is too faithful to the training set and therefore not able to generalize on the test set. Nevertheless, it is possible to

observe that, in the cases in which the aforementioned characteristics are present (e.g., neural network/MLP models with a high number of neurons per layer), excellent results are obtained on the test set and therefore we can conclude that no overfitting occurred. Then, for the sake of completeness, we briefly report on the data preprocessing steps that were required with the different learning platforms. More in detail, we implemented the following procedures: (i) Locate and delete rows containing null values in the dataset. In particular, this operation was expressively needed with Knime and Python. Such a situation is, instead, handled automatically in Weka, Orange, and Ludwig; (ii) Convert the values of the target feature *classe* from categorical to numeric. This operation was only necessary by utilizing the Python programming language with the Keras libraries. The other adopted learning platforms handled this situation automatically. Please note that the least possible steps were implemented to keep the preprocessing phase as simple as possible: we solely aimed at removing any possible error messages. Any other possible data preprocessing (e.g., data normalization) has not been applied to the initial raw data unless already included in the default settings of the different learning platforms.

Concluding, with the increasing development of application packages like the aforementioned (Weka, Orange, Ludwig, and Knime), it is possible to observe how the complexity of applying machine learning algorithms is shifting more and more from possessing specialist knowledge in data science and algorithm implementation to owning the ability to design systems and collect the right data concerning the data analysis problem that is considered. The aim of this work was to demonstrate that there are simple ways of applying machine learning algorithms to easily create models that may still provide significant results related to a specific problem, without necessarily having specialized knowledge in the field of data science. For this purpose, a classification problem from the field of qualitative activity recognition was considered and solved with the help of the machine learning platforms Weka, Orange, Ludwig, and Knime. A further analysis related to the same classification problem was then carried out utilizing the Python programming language. Important to highlight that, to pursue the aim of this work, the models were put to

good use without setting any unnecessary parameters. Our analysis exhibits how the employed learning algorithms, with the considered raw dataset, can be powerful, but parameter-dependent. Therefore, data-human interfaces and the learning algorithms they implement may certainly represent a useful tool for non-expert data scientists, nevertheless, they should carefully be put to good use. In fact, for a considered problem, and the relative analysis, different knowledge (domain knowledge rather than a technical one) could be sufficient/necessary (or not). In some cases, a specific domain knowledge, in addition to the use of learning tools, is enough to obtain good results (e.g., the case explained in detail in this Chapter). Other times, however, both an in-depth domain and algorithmic knowledge are essential. An example will be given in Chapter 4. In the research work there explained, in-depth domain knowledge is essential to understand how to frame learning models, concerning the socio-historical considered tasks. Nevertheless, without in-depth knowledge of machine and deep learning theory, algorithms, and parameters, it would have been practically impossible to be able to build deep learning models capable of obtaining results like the ones here shown.

Chapter 3

Setting the stage for a machine learning task

Considering a Human Reaching Movement analysis in the Psychological field

3.1 Introduction

Our everyday life is deeply defined by the voluntary actions we execute toward ourselves and toward the world that surrounds us. The way we plan and control our movements has been widely investigated for different motor tasks, to deepen our understanding of which motor strategies individuals adopt to select and execute different goal-oriented actions. In particular, as action features are usually movement-specific, we want to focus on a specific arm movement, namely *reaching*, which allows human beings to act within their peri-personal space by grasping, manipulating, and using objects, as well as to interact with their own bodies and with other people. Performing a reaching movement action requires both pre-planning and on-line control of the desired motor output. Such two mechanisms are settled in distinct brain regions, respectively intervening in either the early or later movement time and appear influenced by different sensorimotor aspects and cognitive processes [94]. Indeed, the role of motor networks might go beyond the action specification that answers to the "how to do it" and contribute

to the simultaneous process of action selection, which addresses the "what to do" issue and chooses among currently available options [95]. It goes without saying that cognitive control is fundamental to the process of action selection, including the ability to inhibit inappropriate or incorrect responses [96]. Rather than a unitary process, inhibition is a multifaceted skill that comprehends sensory, cognitive, behavioral, and motor sub-components [97], such as the ability to stop prepotent motor activities.

Then, performing cognitive operations and motor actions can be considered two faces of the same coin, as they vastly rely on shared mechanisms that allow us to produce appropriate responses with respect to goals and context [98]. All relevant processes specialize with age, with motor and cognitive development being closely connected and inter-related in a dynamic process of exploring and adjusting to the demands of the external physical and social environment [99]. Although cognitive and motor difficulties often co-occur in neurodevelopmental conditions and have been extensively studied as separate processes [100, 101], their common underlying mechanisms are still to be furthered. We strongly believe that an integrated approach will provide a more complete understanding of the interplay between low-level sensorimotor processes and high-level executive functioning. Indeed, executive functions are those top-down processes (i.e., working memory, inhibition, and shifting) that enable people to plan, monitor, and control sensorimotor, socio-affective, and cognitive processes, being fundamental to mental and physical wellbeing [102]. Among these functions, the ability to inhibit automatic and highly probable responses, and let less probable alternatives successfully compete for control of cognition and behaviors, ensures that we are flexible and open to learning from the surrounding environment [103].

In addition, the inhibition of prepotent responses is a well-studied process being affected by disorders such as Attention Deficit and Hyperactivity Disorder (ADHD) [104], which is diagnosed based on inattentiveness, impulsiveness and hyperactivity symptoms [105]. On one hand, at the cognitive level, it is established that people with ADHD, despite the wide variability that characterizes developmental trajectories, are overall impaired in executive functions [106]. People with

ADHD can entail several aspects of cognitive and motor impulsivity, which consists of non-reflective stimulus-driven processes and manifests itself through inhibitory difficulties, distractibility, faster, and less accurate responses to neuropsychological tests [107]. On the other hand, at the motor level, it is still debated whether motor signs of atypical development can be detected from infancy and interpreted as early risk factors for the following development of ADHD cognitive and behavioral symptoms [108]. Some co-occurrent difficulties in motor skills (e.g., fine motor precision, manual dexterity, bilateral coordination, balance and postural control, running speed and agility, limb coordination, strength) can be found in about 50% of individuals with ADHD [109]. However, those are not diagnostic criteria and there is no evidence so far supporting the link between motor impairments and ADHD-specific symptoms such as inhibitory deficiencies [109]. To shed light on this, an approach that studies these two aspects in an integrated manner could provide an innovative perspective on difficulties with inhibition and behavioral hyperactivity. Potential underlying mechanisms of inhibition difficulties relate to motor planning, which is responsible for selecting the action target and the timing of movements (e.g., reaction times, movement times, and acceleration/velocity parameters) [94].

In this analysis, the stage set in the present work wants to explore the distinctive contribution of motor planning and control to human reaching movements in the psychological field, in order to set the stage for a possible machine learning task. In particular, the movements were triggered by the selection of a prepotent response (Dominant) or, instead, by the inhibition of the prepotent response, which required the selection of an alternative one (Non-dominant). To this end, we adapted a *Go/No-Go* task [110] to investigate both the dominant and non-dominant movements, firstly, of a cohort of 19 neurotypical adults and, secondly, of a cohort of 17 children with ADHD and 26 children with Typical Development (TD), utilizing different kinematic measures, in order to discriminate between the planning and control components of the two actions. In this analysis, a portable and low-cost 3-axis wrist-worn accelerometer was put to good use to obtain raw acceleration data and compute and break down its velocity components, with the

aim to discuss the possibility to consider an accelerometer-based analysis in a clinical context.

In the first study with neurotypical adults, the obtained results indicate that, with the inhibition of a prepotent response, the selection and execution of the alternative one yields both a longer Reaction Time (RT) and Movement Duration (MD). The Time to Peak Velocity (TPV) appeared higher in the non-dominant response with respect to the dominant response, revealing that participants tended to indulge more in motor planning than in adjusting their movement along the way. Moreover, comparing such results to the findings obtained by other means in the literature, we could verify the feasibility of an accelerometer-based analysis to disentangle distinctive cognitive mechanisms of human movements. The entire analysis and the relative results regarding this first study are published in [3].

In a second study, instead, we aimed to explore the different strategies used by children with ADHD and TD to provide a prepotent response or inhibit the prepotent and select an alternative one, following the same way adopted for neurotypical adults. We hypothesized that children with ADHD, compared to neurotypical controls, would show greater difficulties inhibiting the prepotent response, which the literature also refers to as motor impulsivity [111]. In addition, we expected children with ADHD to make more errors than controls in the non-dominant condition, and show an atypical motor profile, with reduced or less effective motor planning. In particular, as markers of motor impulsivity, we expected reduced RT and TPV in the group of children with ADHD [111]. Nevertheless, although no group difference emerged on accuracy levels, the kinematic analysis of correct responses revealed that, unlike neurotypical children, those with ADHD did not show increased motor planning in non-dominant compared to the dominant trials. In our implemented task, motor control could have compensated leading to good accuracy. However, this strategy might make inhibition harder in more naturalistic situations that involve complex actions. Combining cognitive and kinematic measures could be a potential innovative method for assessment and intervention of subtle differences in executive processes such as inhibition, going deeper with respect to just relying on behavioral outcomes alone. The idea for this second

study was presented in [4, 5], while the entire analysis and relative results relative are published in [6].

The rest of this Chapter is organized as follows. In Section 3.2 previous analyses presented in the literature, regarding the task considered in our analysis, are reviewed. In Section 3.3 the main contributions of these studies are reported. In Section 3.4 the entire experimental setup and procedure are furnished, starting from the recruitment phase (Section 3.4.1) and the participant characteristics (Section 3.4.2), through the procedure and the task details (Sections 3.4.3 and 3.4.4, respectively), until the apparatus (Section 3.4.5). Then, in Section 3.5 the dataset collected is introduced and explained. Follows Section 3.6, in which we present the method exploited for these studies, dividing the acceleration calibration and preprocessing (Section 3.6.1), the velocity and TPV computation (Section 3.6.2), and the statistical approach and analysis (Section 3.6.3). After this, all the obtained results are reported in Section 3.7. In particular, firstly, we report the results regarding the study with neurotypical adults (Section 3.7.1) and, secondly, the ones regarding the study with children with ADHD and TD (Section 3.7.2). The same schema is replicated both for the discussion of the presented results and the conclusions drawn from this work in Section 3.8 (divided in Section 3.8.1 and Section 3.8.2) and in Section 3.9 (divided in Section 3.9.1 and Section 3.9.2), respectively. Successively, in Section 3.10 possible future works are provided.

3.2 Related works

Previous studies were mainly based on correlational analysis of motor skills and purely cognitive performance at inhibition tasks and failed to find clear relationships [109]. In addition, investigating inhibition without dissociating motor and cognitive aspects that are deeply interrelated offers further insights into the underlying processes. Nevertheless, the compelling possibility of integrating a kinematic measure into the traditional neuropsychological evaluation is strongly limited by the need for sophisticated motion capture systems. In fact, those used for research purposes are often expensive and bulky, thus being hardly affordable for most

clinical centers. In order to use low-cost portable solutions and boost the applicability of motion analysis, inertial sensors have been recently recommended for their good measurement reliability and validity [112]. Adopting this technology in clinical practice would allow for a more detailed analysis of the mechanisms underlying the child's performance on tests of interest: it could be used during assessment for setting specific intervention goals, for monitoring treatment effects, and as a treatment tool itself when used as biofeedback.

Different paradigms are commonly used to measure the inhibition of prepotent responses (e.g., Stroop, Stop-signal, and Go/No-Go tasks), with diverse versions that rely on mainly cognitive processes or entail varying degrees of motor components and activate both distinct and shared neural areas [113, 114]. In neuropsychology, one of the most used is the *Go/No-Go* one [110]. On one hand, the "Go" trials require participants to provide a fast response (i.e., do something) as soon as a dominant cue appears. On the other hand, the "No-Go" trials require to inhibit the response and not answer (i.e., do nothing) when another non-dominant cue appears (the latter usually appears less frequently than the dominant one) [115]. The motor component comes into play when the response requires some sort of movement, from pressing a button to reaching a target, which sometimes has to be voluntarily stopped before or during its execution [113]. However, the classical task is unable to investigate the different motor strategies individuals may adopt to perform either a prepotent or alternative response. Then, since the very planning of this motor response could reveal important information about the processes at play, a deeper understanding of motor responses in cognitive tasks needs an improved consideration, leading to a new perspective on the shared mechanisms that underpin adaptive behaviors. In particular, to further distinguish between planning and control aspects in the various phases of action, kinematic measures have been included with adapted a *Go/No-Go* paradigm that asked participants to perform either a prepotent action elicited most of the time (Dominant) or an alternative less frequent one (Non-dominant).

In a possible adaptation of the *Go/No-Go* task, both Reaction Time (RT) and Movement Duration (MD) were calculated and analysed [116]. Researchers

often use RT to indicate the time from the appearance of the "Go" stimulus to the moment when the person gives the response (which corresponds to the end of the movement). In addition, this is the index of choice for studying variability in the inhibitory abilities of people with ADHD [117]. However, using RT as the total response time does not consider the two underlying processes separately: the preparatory activities that take place before the start of the movement and the actual motor execution [118]. Then, we hereby calculate RT as the time from the appearance of the "Go" stimulus to the beginning of the movement, so that it gives us a measure of pure motor planning. A higher need for motor planning is expected to result in higher RTs [119]. The MD, instead, is calculated as the movement execution time from when the response movement begins to when it ends and, across MD, motor planning gradually gives way to control and monitoring of the ongoing movement, which involve distinct processes [94]. Nevertheless, it is worth noting that motor planning is not relegated to RT, but also overlaps with motor control during the MD. Indeed, "as planning is generally operative early and control late in a movement, the influence of each will rise and fall as the movement unfolds" [94, p. 5]. Therefore, kinematic indices other than RT and MD would be more informative to further clarify the mechanisms beneath distinct movements, with promising possibilities to distinguish the specific inhibitory impairments that are common in several neuropsychological conditions [120]. As planning seems to be primarily devoted to processing cognitive information, whereas control is dedicated to homing in on a target with specific spatial features [94], the inhibition of prepotent motor responses evoked by *Go/No-Go* tasks would likely load on planning mechanisms.

The movement research field has extensively debated the distinctive meaning of different motor indices, which are affected by different factors, thus providing insights into distinct neuropsychological mechanisms underlying motor activities. In particular, acceleration discloses the movement smoothness, whereby an optimal reaching is ideal, for instance in experimental contexts and robotics, the one with the minimum jerk, namely the rate of acceleration change in time [121, 122]. The smoothness of a reach-to-grasp movement might depend on whether the tar-

get object is present, imagined, or absent, on how it is oriented, or on which is the plane of movement (e.g., horizontal or vertical plane) [123]. Neuro-imaging studies collected evidence of distinct cortical networks being related to distinct kinematic features. A research work [124] studied the fast repetitive voluntary hand movements of neurotypical adults revealing that movement acceleration was mainly coupled with a coherent activation of contralateral primary motor (M1) hand area at $\approx$ 3 Hz and $\approx$ 6 Hz of movement frequencies. Moreover, only when the hand movement is aimed at touching its own fingers, the primary somato-sensory (S1) hand area became the most coherent brain area at $\approx$ 3 Hz of motion frequency. In addition, the activation of DLPFC area (dorsolateral prefrontal cortex), which is responsible for goal-directed action planning, and PPC area (posterior parietal cortex), which is responsible for sensorimotor integration and movement monitoring, were coherent with movement acceleration [125]. Instead, focusing on velocity, the minimum-jerk model predicts that reaching trajectories starting and ending at full rest will show a symmetric, bell-shaped velocity path, with 50% of MD spent both accelerating and decelerating. However, MD and velocity across time are shaped by several factors, such as the individual developmental trajectory [126], the affordances of the target object (e.g., a cup or a spoon) [123], and social intentions during interactions with others [127]. On this matter, the percent Time to Peak Velocity percentage (TPV%) may represent a useful index to disentangle how much of the movement time is devoted to planning or control. Theoretical (e.g., in robotics) reaching trajectories starting and ending at full rest will show a bell-shaped velocity path, with the first half of MD spent accelerating and the second one decelerating, resulting in a 50% TPV [122, 121]. On one hand, given that whether a kinematic parameter occurs earlier or later over the MD would reflect more either planning or control [94], a small TPV% resulting in a longer deceleration phase may indicate a greater need for control and adjustment of the ongoing movement. On the other hand, a big TPV% resulting in a shorter deceleration phase may indicate a greater need for motor planning.

Regarding people with ADHD, adults have been found to show atypical motor profiles, with longer RTs to start moving after a "Go" cue and higher variability in

the velocity shape over time, suggesting impaired motor planning capacities [128]. It is interesting to note that there is a kind of slowness in sensorimotor and cognitive processes that underlie behavioral manifestations of impulsivity, hyperactivity, and inattention. A developmental perspective is needed to understand how these atypicalities have emerged and are maintained from childhood to adulthood. This would help us design targeted and age-appropriate interventions to promote a change in the mechanisms underlying the cognitive and behavioral difficulties of ADHD. Notably, purely cognitive training specifically targeting executive functions such as working memory, attention, inhibition, and shifting rarely results in cognitive nor behavioral or academic improvements, with scarce effect on ADHD core symptoms [129, 130]. It has been speculated that leveraging embodied cognition and cognitive-motor approaches could boost training efficacy [131]. This multidimensional perspective would eventually chart the way to define and test both motor and cognitive interventions to strengthen inhibition by passing through multidimensional doorways. Despite their presence and impact, motor difficulties of people with ADHD often end up being overlooked by research and clinical practice.

Concluding, the present study aims at setting the stage for a future machine learning task considering a human reaching movement analysis in the psychological field, investigating the human ability to inhibit prepotent motor responses, through an adapted version of the *Go/No-Go* paradigm, implemented from scratch. Neurotypical adults, children with ADHD and Typical Development (TD) were recruited. A commercially available, low-cost, easy-to-use, wearable accelerometer sensor was employed to capture movement features, and the applicability of such a low-cost portable tracking tool in a clinical context was discussed.

3.3 Contributions

Starting from scratch, from an experimental perspective, the main contributions of this study amounts to [3]: (i) Adapt a *Go/No-Go* task to assess the motor planning and control in the selection or inhibition of a prepotent response through kinematics measures; (ii) Implement the aforementioned task to make it accessible

for testing through low-cost components, i.e., a computer with a touch screen and a position sensor; (iii) Collect kinematics data, i.e., raw acceleration data, through a low-cost wearable device, i.e., a 3-axis wrist-worn accelerometer; (iv) Build and validate an entire pipeline of analysis in order to use this low-cost portable motion tracking tool, to boost the applicability of our methods of analysis to a broad range of research and clinical contexts, such as the human reaching movement analysis in the psychological field; (v) Set both the data collection and data analysis to possibly integrate with machine learning algorithms, as soon as the dataset size is enough, in order to increase the applicability of such a pipeline in such a context.

Importantly, this work provides different contributions to the results of human reaching movement analysis in the psychological field. From this perspective, the main contributions amount to: (i) Investigate both the dominant and non-dominant movements of a cohort of 19 neurotypical adults, utilizing kinematic measures to discriminate between the planning and control components of the two actions through a customized low-cost portable motion tracking tool [3]; (ii) Discuss and verify the feasibility of an accelerometer-based analysis to disentangle distinctive cognitive mechanisms of human movements [3]; (iii) Investigate children's ability to inhibit prepotent motor responses, through the same paradigm and apparatus, with a cohort of 17 children with Attention Deficit and Hyperactivity Disorder (ADHD) and 26 children with Typical Development (TD) as the control group, focusing on measuring the velocity shape across movement time ("when") [4, 5, 6].

3.4 Experimental setup and procedure

3.4.1 Recruitment

Neurotypical adults The recruitment took place among university students with no past or present history of clinical conditions (self-reported). The participants voluntarily participated in the study and did not receive compensation.

Children with TD and ADHD For the study with children participants, data collection was planned to take place between December 2019 and April 2020, as part of a collaborative project with a clinical center in northern Italy, which is specialized in ADHD diagnosis and intervention. Later, however, data collection was interrupted at the beginning of the Covid-19 pandemic and resumed when the center was authorized to reopen to external operators (i.e., the investigators). Thus, a further phase of data collection was carried out between October and December 2021. The partner center had an average intake of 60 children, and all were offered voluntary participation in the study. The final sample of children with ADHD was determined by the number of parents and children who joined and participated. Since ADHD is an inherently heterogeneous condition [132, 133], we have not established inclusion or exclusion criteria based on IQ, level of support needed, or possible presence of co-occurring medical or neuropsychological conditions. Thus, we aimed to include participants from the heterogeneous ADHD population. Psychologists confirmed children's diagnoses and provided IQ assessments through the WISC-IV scale. Moreover, we collected parent-reported questionnaires on the child's executive (Executive Functions Questionnaire - Q.FE [134]) and sensory profile (Short Sensory Profile - SSP [135]), as well as the presence and severity of restricted and repetitive behaviors (Repetitive Behavior Scale-Revised - RBS-R [136]). A convenient control group of children with Typical Development (TD) in the same age range was tested at the University of Padova. According to parents' reports, typically developing children had no medical or neuropsychological conditions.

3.4.2 Participants

Neurotypical adults For the first study, which also includes the discussion about the feasibility of an accelerometer-based analysis to disentangle distinctive cognitive mechanisms of human movements, we recruited 19 neurotypical adults aged from 18 to 26 years old (M = 22.3, SD = 1.9), among them 5 men.

	IQ	RBS Tot	Low-level RRB	High-level RRB	Q-FE Tot	SSP Tot
M	107.2	21.8	7.8	14.0	90.6	136.4
SD	17.6	15.5	6.5	11.8	20.9	26.2

Table 3.1: ADHD group characterisation ($n_{ADHD} = 17$). Mean (M) and Standard Deviation (SD) of: *IQ*, total score from the WISC-IV scale; *RBS Tot*, total score from the RBS-R. Higher scores indicate a more severe profile of restricted and repetitive behaviors; *Low-level RRB*, scores from Stereotyped, and Self-Injurious subscales of the RBS-R; *High-level RRB*, scores from Compulsive, Ritualistic, Sameness and Restricted Interests behaviors subscales of the RBS-R; *SSP Tot*, total score from the SSP. Higher scores indicate better sensory profile; *Q-FE Tot*, total score from Q-FE. Higher scores indicate better executive functions.

Children with TD and ADHD For the second study, instead, we recruited 17 children with ADHD (4 female children) from 6 to 15 years of age (M = 9.4, SD = 2.2), and 26 children with Typical Development as control group (10 female children), from 6 to 13 years of age (M = 9.2, SD = 2.1). Three additional participants (2 in the ADHD and 1 in the TD group) were excluded due to technical issues that prevented them from completing at least 50% of the trials of the task. Characteristics of the ADHD group are provided in Table 3.1, which includes IQs and scores from the parent-reported assessment. A total of 12 children were diagnosed with the combined subtype of ADHD, 2 with the inattentive subtype, and 3 with the impulsive/hyperactive subtype. Moreover, 6 children received a comorbid diagnosis of specific reading disorders (from moderate to severe), 2 children received a diagnosis of a specific spelling disorder (moderate), and 4 were diagnosed with other behavioral and emotional disorders.

3.4.3 Procedure

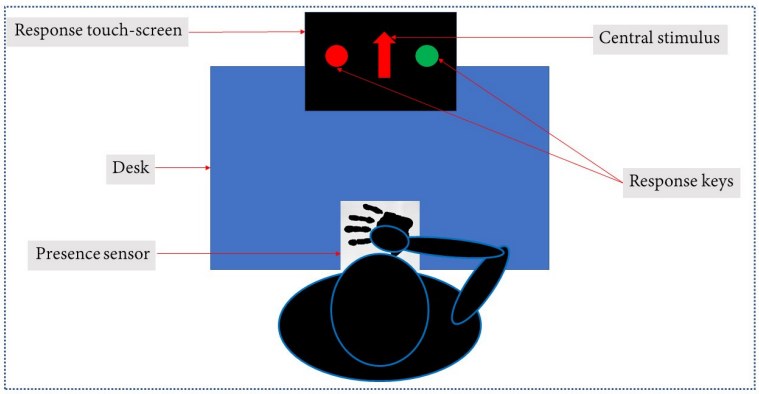
Participants were welcomed into the lab and asked to sign a written consent form (for the children participants, their parents signed it). The study and all the

experimental methods were approved by the Research Ethics Committee of the School of Psychology, University of Padova (protocol no. 3251). The experiment was carried out in accordance with the approved guidelines and regulations.

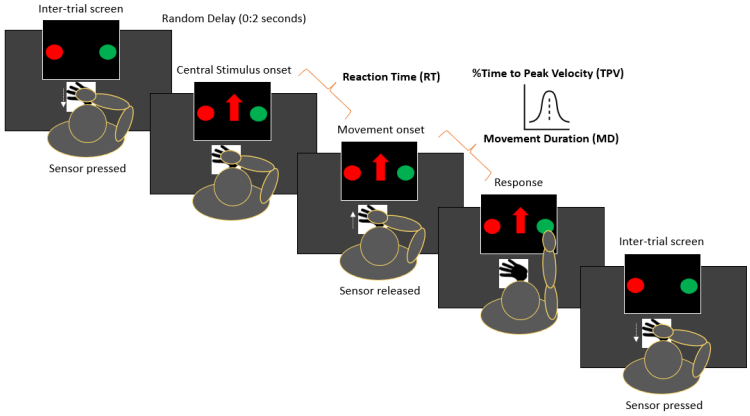
The participants sat at a desk and wore an accelerometer research watch on their dominant wrist (a schema of the experimental setup is reported in Fig. 3.1a). They were instructed to place the dominant hand at a specific starting position, monitored by a presence sensor. At the distance of their arm length, they found a response touchscreen, so they were required to completely extend their arm to touch the response screen. A specific task (details in Section 3.4.4) was proposed and required the participant to make action selection choices by touching one of the response keys on the screen. Upon comparison of a central stimulus, participants were asked to select, reach, and press one of two response keys placed one on the left and one on the right side of the central stimulus, following specific instructions. Before the start of the next trial, participants had to return their hand to the sensor. As soon as the hand was in place, the next trial started after a random delay (in a range from 0 to 2,000 ms), which prevented participants from anticipating the onset of the next trial. The task tested the participant's ability to select a prepotent or an alternative response (a schema of the experimental procedure is reported in Fig. 3.1b) and, during this behavioral task, the kinematics of the participant's dominant arm was monitored by the wrist-worn 3-axis accelerometer.

3.4.4 Task

The task protagonist in this study focuses on the motor planning and control in the *selection or inhibition of a prepotent response* and a *Go/No-Go* paradigm was adapted to assess this phenomenon. More in detail, upon a comparison of a central stimulus (red/green, upwards/downwards arrow), participants were asked to select, reach, and press one of two response keys (either a red circle or a green circle) placed one on the left and one on the right side of the central stimulus, following specific instructions. Participants were told to select (a) the response key of the same color of the central stimulus when it was an upwards/downwards (counterbalanced between participants) arrow (dominant condition) and (b) the response



(a)



(b)

Figure 3.1: Experimental (a) setup and (b) procedure.

key of the different color when the central stimulus was an averted (either upwards or downwards, counterbalanced between participants) arrow (non-dominant condition). We built a prepotent response for the same-color action, given that it was the one that appeared with a higher chance (75%). On the contrary, we elicited an inhibitory different-color action, which was the less probable one (25%). In this way, we were able to measure the kinematics (details in Section 3.6) of dominant vs. non-dominant selections, being the movements equal.

Participants were instructed to reply as quickly and accurately as possible. Failure to press any keys within $2,000$ ms was marked as "omission". When partic-

ipants moved their dominant hand from the starting position before the appearance of the cue stimulus, instead, the response was tagged as "anticipation" and the program aborted the trial by providing no cue stimulus. Omissions and anticipations were considered invalid trials, therefore excluded from the analysis. For this task, each participant was required to perform 160 valid trials (i.e., trials with correct/incorrect answers). In any case, the total trials never exceed a maximum number of 180, if omissions and anticipations occurred. In addition, two blocks of trials were administered, distinguished by the red/green response keys being located once on the right and once on the left side of the touchscreen. To maintain participants' engagement during the task, a short (30 seconds on average) video from well-known movies appeared every 40 trials. As aforementioned in Section 3.4.3, before the start of the next trial, the participant had to return his hand to the sensor, which prevented participants from anticipating the onset of the next trial. As soon as the hand was in place, as long as the previous trial was not running anymore, the next trial started after a random delay in the range from 0 to $2,000$ ms. We will refer to this independent variable as *StimulusRandomTime* and analyse its effect on participants' performance. Indeed, this variable manipulated the time available to pre-activate the sensorimotor system and predict the incoming occurrence of the central stimulus, potentially affecting the response timing [137]. A schema of the experimental setup and procedure is reported in Fig.3.1. The task lasted about 15 minutes.

3.4.5 Apparatus

Although motor analysis is highly informative both in research and clinical settings, kinematic studies often rely on expensive, bulky, and sophisticated motion capture systems which may not be affordable in most operative and experimental contexts. To use low-cost portable solutions and boost the applicability of motion analysis, both custom-made [138] and commercial tools have been recently evaluated. One extensively used commercial option is the Leap Motion Controller system, a small compact device containing two cameras and three infrared light diodes which have, however, spatial and temporal limits compared to motion cap-

ture systems [139]. Another commercial possibility that seems more promising in terms of measurement reliability and validity is the inertial sensors built with 3-axis accelerometers, gyroscopes, and magnetometers. In particular, Cahill-Rowley and Rose [112] analyzed human reaching kinematics through both inertial sensors and gold-standard motion capture systems. The two methods provided consistent measures of displacement, peak velocity magnitude and timing. In light of this encouraging evidence, the time is ripe for the use of low-cost accelerometers to investigate distinct neuropsychological mechanisms beneath action selection. In particular, in the present study:

- We employed the GENEActiv Original 3-axis wrist-worn accelerometer [16] (size: 43 mm $\times$ 40 mm $\times$ 13 mm, weight without the strap: 16 g) to monitor participants' arm movements. The device measured accelerations through a MEMS sensor, within a range of $\pm$8 g, at a 12 bit (3.9 mg) resolution with a 100 Hz logging frequency;

- The task was implemented resorting to a JavaFX-based application [140];

- To run the experiment, we employed a laptop Lenovo G50-80 (Intel Core i5-5200U (2.2 GHz), 4 GB DDR3L SDRAM, 500 GB HDD, 15.6" HD LED (1366 $\times$ 768), Intel HD Graphics 5500, Windows 10 64-bit);

- The analysis of the resulting data was performed resorting to Python [83] and primarily to pandas, NumPy, and SciPy libraries;

- Participants responded by tapping on a 19 inch touchscreen (LG-T1910BP), with response time 5 ms;

- The presence-absence of the participant's hand in the starting position was detected through a custom-made presence sensor based on Arduino Leonardo which sent the hand detection data to the laptop via one of its USB ports. The presence sensor was connected to a ground capacitor (100 pF) and a capacitive sensor, which consisted of a copper foil wrapped with plastic film (dimension 20 cm $\times$ 12 cm, thickness 0.1 mm) and the presence sensor program was written using the Arduino Capacitive Sensing Library.

3.5 Dataset

Following the experimental setup and procedure described in Section 3.4, different information is collected in order to proceed with the analysis. In particular, for each participant (both adults and children participants) are reported a number, instead of the proper name (for a privacy constraint), *Gender*, and *Age*, as demographic information and the *Group*, i.e., neurotypical adults, children with TD, or children with ADHD. In addition, regarding the children with ADHD, other information is collected. Some characteristics of the ADHD group are reported, in a cumulative format, in Table 3.1, Section 3.4.2. Nevertheless, this information is not directly used for the analysis here reported. Instead, for each trial, are reported: the *Condition* of the trial, i.e., dominant or non-dominant; the raw acceleration values along the 3-axis, i.e., x, y, and z, of the wrist-worn accelerometer; the *Trial Evaluation*, i.e., valid, anticipation, or omission; the *Answer Evaluation*, i.e., correct or incorrect; the time instant in which the sensor is pressed (P), i.e., before the trial start; the time instant in which the central stimulus appears on the touch screen (S), i.e., when the trial starts; the time instant in which the sensor is released (R), i.e., when the answering movement starts; the time instant in which a button on the touch screen is clicked (A), i.e., when an answer is given. In this way, we are able to select valid trials to analyze and compute different measures, among which the Reaction Time (RT), as the time passed from S to R, and the Movement Duration (MD), as the time passed from R to A. In addition, as better explained in the following Section 3.6, we also considered the independent variable *StimulusRandomTime* as the time passed from P to S, randomly set in the range from 0 to $2,000$ ms, in each trial. In total, the dataset comprehends $3,145$ trials (i.e., rows) from the neurotypical adult participants, $4,526$ rows from the children with TD, and $3,023$ from the children with ADHD [141, 6].

Importantly, collecting data following this procedure gives us the possibility to apply machine learning algorithms to the same task, or to another based on the same dataset, once we have reached a sufficient quantity of data for this class of algorithms.

3.6 Method

During the experimental phase, different data was collected both regarding the task application and the movement kinematics measures, as just described in Section 3.5. In this Section, we will focus on the method to follow in to analyze them. Initially, for each valid trial (i.e., no anticipation, no omission) we reported different time instants: sensor pressed, stimulus appeared, sensor released, and answer given, which from now on we will refer to as P, S, R, and A, respectively. To obtain these data, we synchronized the software logs and the accelerometer with the computer local time, thus combining the accelerometer data with the task outputs. The time intervals that are related to the kinematic measures of interest were [S,R], which defined RT, and [R,A], which corresponded to the MD and was used to compute the TPV%. In addition, the interval [P,S] determined the *StimulusRandomTime*. Then, we started with the computation of the effective acceleration, individuated through raw accelerometer data calibration and preprocessing. Subsequently, we computed velocity and the Time to Peak Velocity percentage (TPV%), which is the percentage of time spent from R to maximum peak velocity in the time interval from R to A (i.e., the MD). From a theoretical and mathematical point of view, the most direct way to start computing the TPV% is by applying an integration in time and obtaining velocity from acceleration. In particular, let $a(t)$ be the acceleration signal on one axis, the related velocity signal $v(t)$ can be computed as $v(t) = \int_{t_i}^{t_f} a(dt)dt + C$, where t_i and t_f are the initial and final time instants of the movement and C is an integration constant. However, when faced with real data and numerical functions (e.g., numerical integration), numerical errors can return unreliable velocity values. In the following, we walk through the methodology adopted to compute the TPV% value starting from the accelerometer data collected during the experimental phase.

Important to note that the acceleration calibration and preprocessing analysis have been run on the data collected by an external experimenter (not part of the cohort involved in the trials) who repeated multiple selection tasks, just as a participant. Within the task, the experimenter answered a central cue stimulus by tapping a central response key below the cue. In this way, the displacement re-

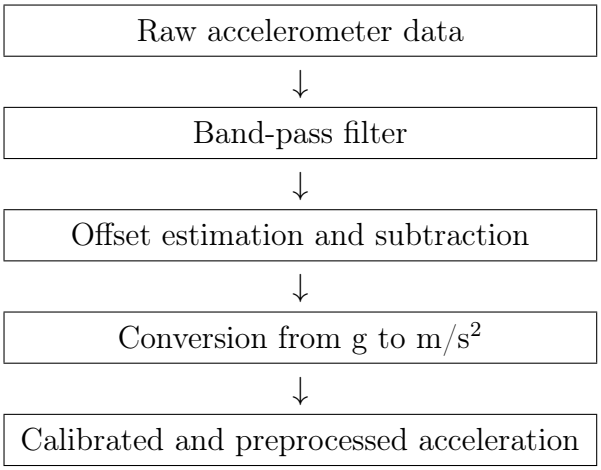
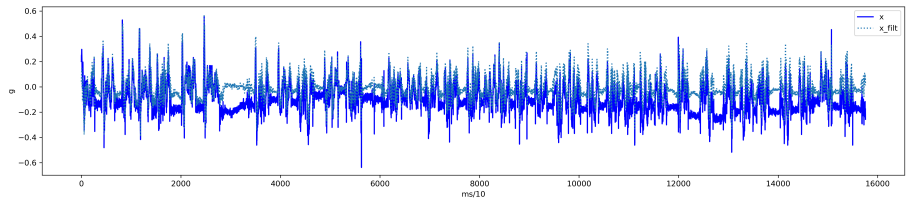


Figure 3.2: Schema of acceleration calibration and preprocessing.

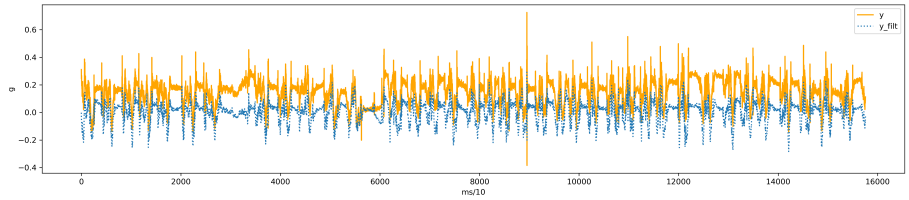
mained roughly the same for each trial. In particular, the experimenter performed 40 trials: 1 anticipation, 2 omissions, and 37 valid answers. The subsequent analysis focused on the raw acceleration signals that started when the sensor was pressed for the first trial and ended when the last valid answer was given.

3.6.1 Acceleration calibration and preprocessing

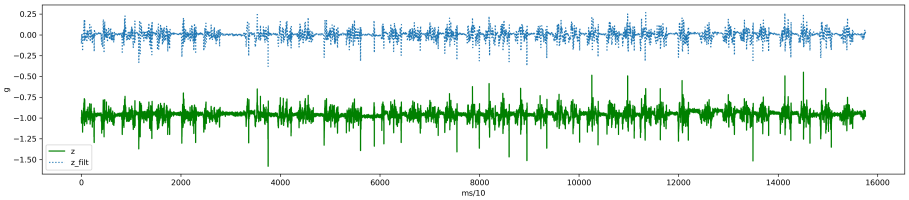
The accelerometer data were sampled at 100 Hz (i.e., data sampled every 10 ms) and stored in g units for offline analyses. Considering the 3-axis accelerometer, the principal output was, for each axis, the measured signal, *acquired acceleration*, which may be broken into *effective acceleration*, *gravity acceleration*, and *noise* [142]. To examine the true movements of the participants, we processed the *acquired acceleration* components to obtain their corresponding *effective acceleration* ones, as raw acceleration signals also contained noise, which could include an offset error, and gravity. In particular, the separation of the latter components becomes increasingly difficult during rotational movements. In fact, in the case of rotational movements (which were observed during our experimental task), the frequency domains of the movement-related component and the gravitational component can overlap, thus their separation can become challenging [142]. Resorting to the state-of-the-art approaches [142], the effective acceleration was extracted by



x component



y component



z component

Figure 3.3: Acceleration signals before (x, y, z) and after (x_filt, y_filt, z_filt) the band-pass filter application.

implementing the following two key steps: (a) a band-pass filter, and (b) an offset estimation and subtraction step. A schema is reported in Fig. 3.2. Following [142], a 4th order Butterworth band-pass filter with cut-off frequencies equal to 0.2-15 Hz was applied to the signal. The filter cut-off frequency of 0.2 Hz was chosen on the presumption that most daily movements of human body parts occur at frequencies higher than 0.2 Hz. The cut-off frequency of 15 Hz was, instead, chosen to remove the effect of high-frequency noise. Also, the 1-20 Hz cut-off frequencies were evaluated, considering other choices made in literature [125, 142, 122, 143], but it was possible to observe no meaningful difference for the 0.2-15 Hz band.

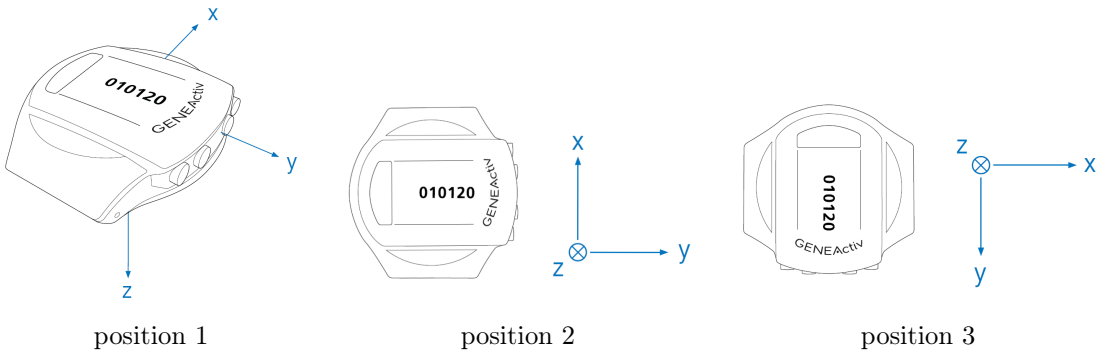
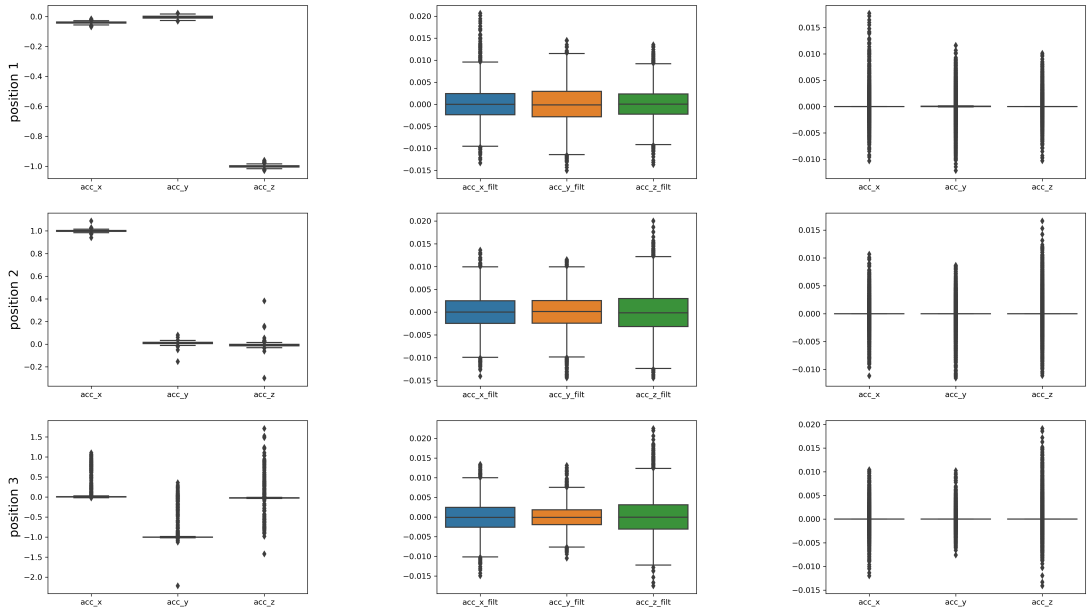


Figure 3.4: Accelerometer at rest positions [16].

Algorithm 1 Accelerometer offset

procedure (for each axis x, y, z)
  i in (x, y, z)
  $df\_filt\_acc \leftarrow$ DataFrame with filtered acceleration
  $df\_offset\_acc \leftarrow$ new DataFrame for offset acceleration
  $epsilon_i \leftarrow$ offset value for axis i
  for j in range (0, len(df_filt_acc)): **do**
    if $df\_filt\_acc[j, acc_i] < (epsilon_i \cdot (-1))$ **then**
      $df\_offset\_acc[j, acc_i] = df\_filt\_acc[j, acc_i] + epsilon_i$
    else if $df\_filt\_acc[j, acc_i] > epsilon_i$ **then**
      $df\_offset\_acc[j, acc_i] = df\_filt\_acc[j, acc_i] - epsilon_i$
    else
      $df\_offset\_acc[j, acc_i] = 0$
    end if
  end for
end procedure

Comparing now the acceleration signals (x, y, and z, in Fig. 3.3) it is possible to see that the raw acceleration components were shifted for 0 g due to the gravity. The z component, for example, would fall as low as $-$g. After applying the band-pass filter, all acceleration components adjusted to lie around 0 g (x_filt, y_filt, and z_filt in Fig. 3.3). To estimate the offset error, data was collected from the accelerometer while at rest with the x, y and z axes pointing towards the ground (see Fig. 3.4). From the filtered signal, for each of the three components, we computed the mean of the differences between actual accelerometer readings and the 0 g value expected from an accelerometer at rest. Hence, we obtained an offset value for each of the three axes. Successively, such values were removed



(a) (b) (c)

Figure 3.5: Acceleration values in g sampled at 100 Hz from the accelerometer at rest: (a) no filtering, (b) band-pass filtering, and (c) band-pass filtering and offset removal (n_{data} for each position $= 6,960$).

from the acceleration data components, according to the pseudocode reported in Algorithm 1. The visualisation of the signal from the accelerometer at rest fixed in the three different positions shows the filter effect and the presence of an offset error (Figs. 3.5a and 3.5b). Indeed, the offset removal led to data closer to zero (Fig. 3.5c). Finally, we obtained an estimate of the effective acceleration, adopting g $=$ 9.80665 m/s^2 for the conversion from g to m/s^2 units.

3.6.2 Velocity and TPV computation

Considering now the calibrated and preprocessed acceleration, let acc_{RA} be the signal related to the time interval [R,A] of a specific valid trial, we applied the cumulative trapezoidal numerical integration function in order to compute velocity. In Fig. 3.6, we reported the velocity components obtained by applying this function to the acceleration values of a trial. After this step, we computed the magnitude (which represents the velocity module) from its components, always shown in Fig. 3.6. Nevertheless, the application of an integration function could lead to an incremental numerical error due to a possible bias (i.e., additive noise) present in the acceleration (visible in Fig. 3.6), whereby the x component and magnitude of velocity present increasing monotonous curves rather than the expected bell shape. Such a phenomenon may lead to the creation of a "new" and "false" maximum peak at the end of MD, making the computation of the central "true" peak quite challenging. To overcome this issue, we applied the detrend function to the velocity magnitude, thus removing the signal linear trend and reducing the numerical error described above. Further details are then reported through an additional and more general analysis.

In the following, with no loss of generality with respect to the aims of the procedure here described, we consider the $sin(t)$, $2sin(t)$, and $3sin(t)$ waveforms as exemplar acceleration components signals. Therefore, we proceeded to compute the velocity components integrating the acceleration ones and obtaining the velocity magnitudes, reported in Fig. 3.7a. After that, we applied the detrend function to the velocity magnitude, as shown in Fig. 3.7b. From these results, it is possible to see that the application of detrend function only modified the signal with respect to the ordinate axis, but did not change the signal shape. This is because the velocity magnitude is computed from acceleration components characterized by neither trend nor bias. Repeating the same analysis, but starting from acceleration components where each of these has a constant bias, instead, we obtained the velocity magnitude before and after the application of detrend function, as shown respectively in Figs. 3.7c and 3.7d. In particular, in Fig. 3.7c it is possible to see an incremental numerical error due to the presence of the acceleration bias,

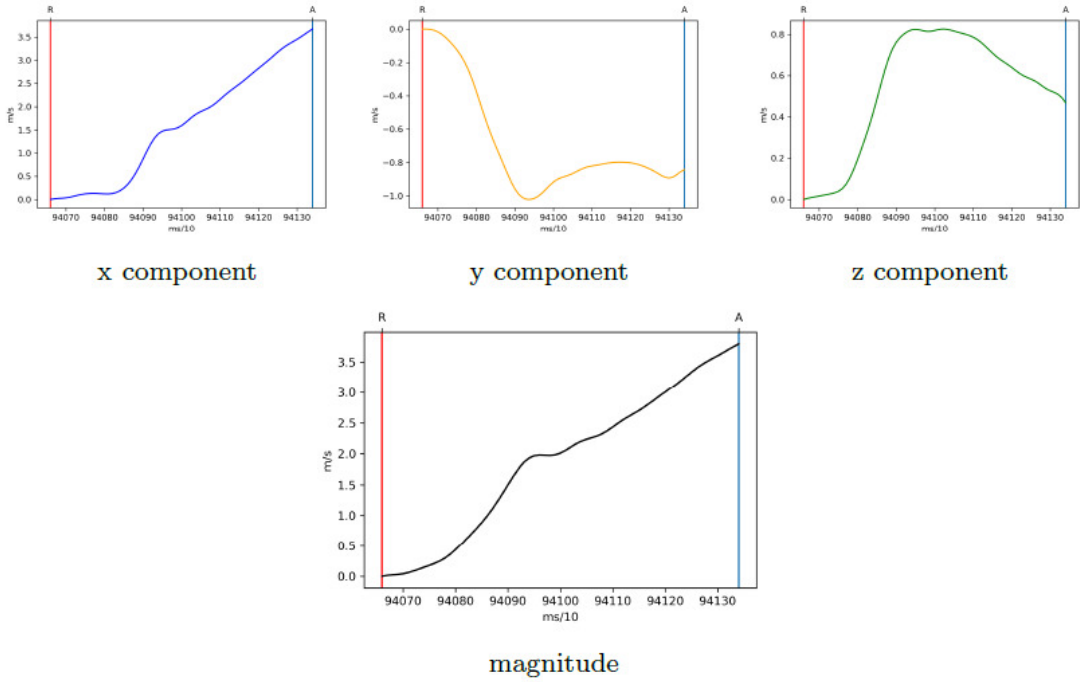


Figure 3.6: Velocity signals of a trial where the error due to the acceleration bias is visible in both the x component and the magnitude (increasing monotonous curves that do not represent the expected bell shape).

as this is amplified by the application of the numerical integration function. Both the signal shape and the signal peak changed. Nevertheless, after the application of the detrend function, some of the signal changes due to this numerical error were removed, as reported in Fig. 3.7d. It is important to note that, comparing velocity signals in Figs. 3.7b and 3.7d, (a) the peak values changed, but (b) the peak position in time is the same. From this exploratory analysis of the signals, hence, it is possible to conclude that, although the velocity values could change due to the detrend function application, the position in time of the peak velocity remains stable. This property meets the requirement of individuating the TPV value set in this work.

As reported in this additional analysis, while the velocity values could change

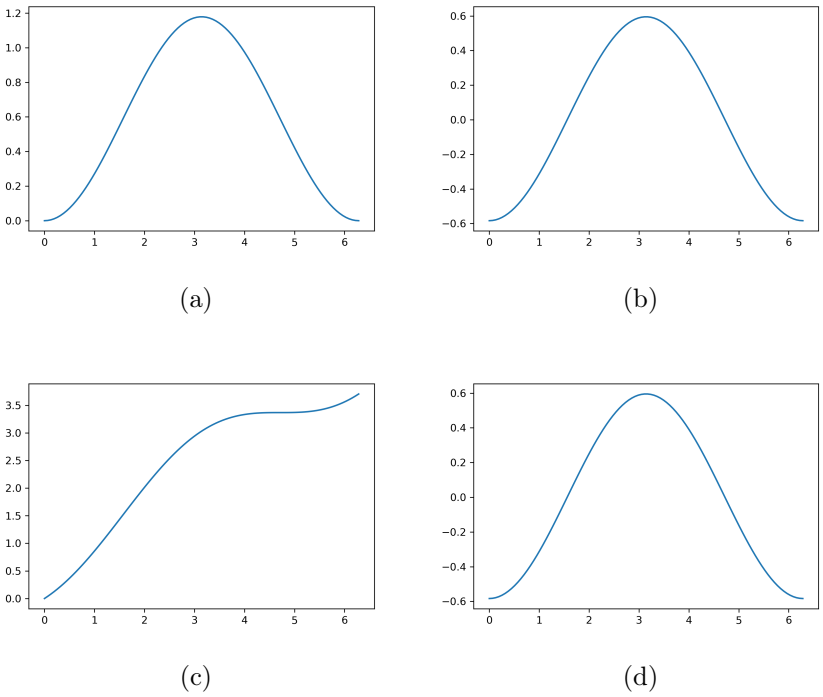


Figure 3.7: Velocity magnitudes obtained from the integration of the acceleration vector components, (a)-(b) with and (c)-(d) without a constant bias, (a)-(c) before and (b)-(d) after applying the detrending.

due to the detrend function application, the position in time of the peak velocity appeared stable, thus allowing us to calculate the TPV% ("when"). Though, we were not able to further investigate those indices based on the velocity value ("how fast"), e.g., the mean velocity, the value of peak velocity. To better clarify this aspect, we adopted a mathematical approach to assess the reliability and validity of the calibrated and preprocessed acceleration and the computed velocity values. As for the calibration and preprocessing analyses (Section 3.6.1), we considered the data collected by an experimenter not belonging to the cohort involved in our trials. We measured the distance between the sensor and where the response keys appear on the touchscreen, corresponding to the actual hand displacement required to reach the screen. Then, we compared such displacement to the one computed from the acceleration data. In particular, we calculated the displacement

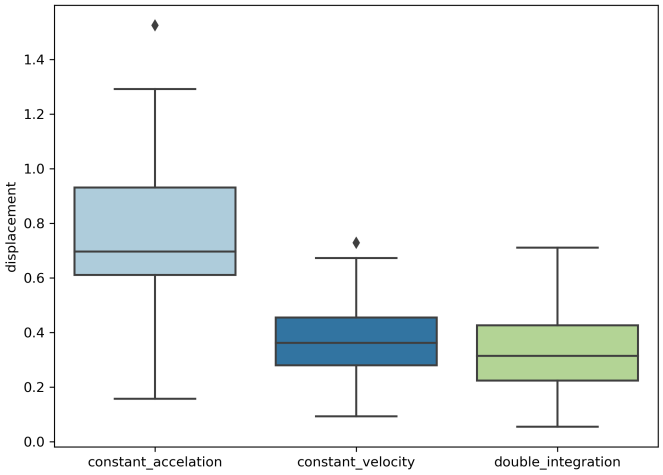


Figure 3.8: Displacement values in m computed from the acceleration values with different methods: constant acceleration, constant velocity, and double integration.

of interest (from R to A, for each valid trial) in three different ways:

(i) Firstly, under the hypothesis of constant acceleration, for each trial, we computed the mean acceleration from R to A, and the displacement as the product between the mean acceleration and the square of time required to cover the distance of interest divided by 2 (i.e., according to the equation of uniformly accelerated motion);

(ii) Secondly, under the hypothesis of constant velocity, for each trial, we computed the mean velocity from R to A, and the displacement as the product between time from R to A and mean velocity (i.e., according to the equation of uniform motion);

(iii) Finally, for each trial, we computed the displacement by applying a double numerical integration to acceleration, using the cumulative trapezoidal numerical function, which does not rely on any hypothesis regarding acceleration or velocity.

Method	M	SD	Actual displacement
Constant acceleration	0.76	0.28	0.46
Constant velocity	0.38	0.15	
Double integration	0.34	0.16	

Table 3.2: Computed and actual displacement values in m ($n_{trials} = 37$).

To calculate the displacement following the aforementioned procedures, the signal was not subject to detrending. The detrending, in fact, can affect velocity component values, which is not acceptable when aiming to compute its magnitude. For this reason, the contribution of the numerical errors may be expected to appear in the displacement. The boxplot of the calculated displacement values is visualized in Fig. 3.8. Then, we computed the Mean (M) and the Standard Deviation (SD) among all trials (37 trials with answer, valid trials). For each method, these results are reported in Table 3.2 and compared to the actual displacement. Notably, the mean values are distant from the actual displacement and the standard deviations are quite high, especially under the hypothesis of constant acceleration. This could be due to the fact that the assumption of neither a constant acceleration nor a constant velocity is appropriate to the actual characteristics of our task. Moreover, a double integration to compute displacements from acceleration can lead to large numerical errors, making this a weak method to assess the reliability and validity of velocity values. Indeed, this computation could be principally impeded by the accumulation of the numerical errors discussed so far. Therefore, with this study, we were not able to confirm nor disprove the reliability and validity of acceleration and velocity values obtained from a setting based on a wrist-worn sensor. Nevertheless, we were able to show that such an approach may be put to good use to obtain the peak velocity timing ("when" in time) information, e.g., time to peak velocity, that may be fruitfully used to analyze the response of a subject.

Ultimately, we aimed to exclude possible extreme TPV% values that would be due to numerical errors, in cases where the detrend function was not sufficient to

remove their effect on the signal. Furthermore, we aimed to remove those TPV% values that were unlikely related to task-related human reaching movements, but rather potentially ascribable to extra-task movements. For these reasons, the a-priori inclusion criteria for valid TPV% values comprehended those values between 5% and 95%.

3.6.3 Statistical approach and analysis

Once the TPV% values were computed, along with RT and MD values, and the answer accuracy observed, we proceeded with different analyses exploiting statistical approaches. In light of the novelty of our paradigm, an exploratory approach was elected to test different potential hypotheses through a model comparison. Importantly, differently from the second study (with children with TD and ADHD), during the first study (with neurotypical adults) the objective was not only to analyze the participants' movements during the adapted *Go/No-Go* task but also to test and validate the experimental procedure, task and apparatus considered, which is based on the use of a low-cost wrist-worn 3-axis accelerometer.

Neurotypical adults For the first study, we investigated whether the TPV% was influenced by the random effect of participants (i.e., interpersonal variability), as well as the fixed effect of *Condition* (within-subjects, two levels categorical factor: dominant vs. non-dominant). Moreover, we checked for the effect of the random time before the central stimulus onset. The latter was a continuous independent variable that we named *StimulusRandomTime*. Each research hypothesis was specified as a statistical model, such that the statistical evidence of the formalised models was evaluated using information criteria [144].

Mixed-effects models were employed to account for the repeated measures design of the experiment (i.e., trials nested within participants). In particular, generalized mixed-effects models were used considering the Beta distribution (with logit link function) of our dependent variable (TPV%). Indeed, the TPV% contained continuous proportions on the interval $(0, 100)$, easily rescaled in the interval $(0, 1)$ (TPV), and can be approximated by a Beta distribution [145]. The statisti-

Model	Dependent variable	Random effect	Fixed effects
`mb0`	TPV	Participants	–
`mb1`	TPV	Participants	*Condition*
`mb2`	TPV	Participants	*Condition* + *StimulusRandomTime*
`mb3`	TPV	Participants	*Condition* × *StimulusRandomTime*

Table 3.3: Model specification for the first study with neurotypical adults.

cal analyses were conducted using the R version 4.0.2 [146], with the glmmTMB package [147] to run the model comparison.

Therefore, we specified four nested models considering the TPV as the dependent variable and the random effect of participants:

- `mb0` (null model) specified the hypothesis of no difference due to the independent variables and only accounted for the random effect of participants;
- `mb1` specified the hypothesis of a difference due to the *Condition* effect;
- `mb2` specified the hypothesis of a difference due to the additive effect of *Condition* and *StimulusRandomTime*;
- `mb3` specified the hypothesis of a difference due to the interaction effect of *Condition* and *StimulusRandomTime*.

In addition, the details of the model specification are depicted in Table 3.3. The four models were compared both through the Akaike weights (i.e., the probability of each model, given the data and the set of considered models) [144], using the R package AICcmodavg [148] and a likelihood ratio test (`anova(mb0, mb1, mb2, mb3)` R function).

Children with TD and ADHD For the second study, we considered 4 dependent variables. An exploratory approach was elected to test different potential hypotheses linking each dependent variable to the predictors of interest.

Through separated sets of model comparisons, different research hypotheses were specified as statistical models, and their statistical evidence was evaluated using information criteria [144]. We separately investigated whether each dependent variable (Accuracy, RT, MD, TPV) was influenced by the fixed effects of *Condition* (within-subjects, two levels categorical factor: dominant vs. non-dominant), *Group* (between-subjects, two levels categorical factor: ADHD vs. TD), and *Age* (continuous numeric variable). All models accounted for the random effect of participants (i.e., interpersonal variability).

Generalized mixed-effects models were employed to account for the repeated measures design of the experiment (i.e., trials nested within participants, which has been included as a random effect in the analyses) and specify the distribution of each dependent variable. For each dependent variable, a set of models were compared through the Akaike weights [144], using the AICcmodavg [144] R package. Then, likelihood ratio tests were used to compare the chosen models and test the effects predicted by the best model. In addition, as an index of goodness of prediction, conditional R^2 (the ratio of variance explained by fixed and random effects over total variance) and marginal R^2 (the ratio of variance explained by fixed effects over total variance) were calculated to quantify the variance explained by the whole model (including the contribution of individual variability) or the fixed effects only (excluding the contribution of individual variability) [149]. Higher percentages of explained variance indicate a stronger strength of association between the dependent variable and the predictors, with the selected model making better predictions. The analyses have been run with R, version 4.0.2 [150].

Therefore, we considered the five models that follow:

- `m0` (null model) specified the hypothesis of no difference due to the independent variables and only accounted for individual variability;
- `m1` specified the hypothesis of a *Condition* effect;
- `m2` specified the hypothesis of additive *Condition* and *Group* effects;
- `m3` specified the hypothesis of additive *Condition*, *Group* and *Age* effects;

- `m4` specified the hypothesis of a two-way interaction effect between *Condition* and *Group*, with the additive *Age* effect.

3.7 Results

3.7.1 Neurotypical adults

The 19 participants provided 2,962 correct responses, 54 incorrect ones (24 in the dominant condition and 30 in the non-dominant condition), 107 omissions (78 in the dominant condition and 29 in the non-dominant condition), and 22 anticipations [3]. As aforementioned in Section 3.6, we only included responses whereby the TPV% was within the 5-95% range, thus considering extremes as due to extra-task movements. At the end of this procedure, we excluded 59 out of 2,962 trials.

RT, MD, and TPV Minimum and maximum values, Means (M) and Standard Deviations (SD) of RT, MD, and TPV% of correct responses in each condition are reported in Table 3.4. The distribution of TPV values in each condition is shown in Fig. 3.9. The model comparison outputs, namely the degree of freedom (Df), Akaike weights ($AICcWt$), chi-squared test statistic values (χ^2), and p-values (p) are reported in Table 3.5.

The most plausible model, given the data and the set of considered models, was `mb2` ($AICcWt = 0.44$), which included the random effect of participants, the additive effects of *Condition* (statistically significant according to $p < .001$), and *StimulusRandomTime* (statistically non-significant according to $p = .08$); p-values from `summary(mb2)` R function. These effects are depicted in Fig. 3.10.

3.7.2 Children with TD and ADHD

Children with TD (26 participants) provided 4,104 valid responses out of 4,526 total trials (91%). Children with ADHD (17 participants) provided 2,472 valid responses out of 3,023 total trials (82%) [6]. This demonstrates both successful

	RT					MD				
Condition	$\mathbf{n_{trials}}$	**min**	**max**	**M**	**SD**	$\mathbf{n_{trials}}$	**min**	**max**	**M**	**SD**
Dominant	2,253	62 ms	1,373 ms	558 ms	136 ms	2,253	266 ms	1,562 ms	500 ms	167 ms
Non-dominant	709	335 ms	1,365 ms	601 ms	163 ms	709	291 ms	1,562 ms	591 ms	207 ms

	TPV%				
Condition	$\mathbf{n_{trials}}$	**min**	**max**	**M**	**SD**
Dominant	2,213	5.04%	94.36%	40%	15%
Non-dominant	690	6.9%	94.31%	45%	17%

Note: TPV% includes less trials due to the exclusion of extreme values

Table 3.4: Neurotypical adults, descriptive statistics ($n_{participants} = 19$).

Model	***Df***	***AICcWt***	χ^2	***p***
`mb0`	3	0.00		
`mb1`	4	0.20	49.70	<0.001
`mb2`	5	0.44	3.58	0.06
`mb3`	6	0.36	1.63	0.20

Table 3.5: Neurotypical adults, model comparison.

task competition (with our task being adequate for both groups) and a low rate of discarded data. As aforementioned in Section 3.6, we only included responses whereby the TPV was within the 5-95% range, thus considering extremes as due to extra-task movements. In addition, for the children participants, that may perform more extra-task movements, we planned to exclude those responses whereby either RT or MD was less than 100 ms, being them ascribable to anticipation. At the end of this procedure, from valid trials performed by both groups, we excluded 217 out of $6,576$ responses (3.3%), and the excluded responses were not further analyzed. Then, the final dataset comprehended $6,359$ observations. In particular, children

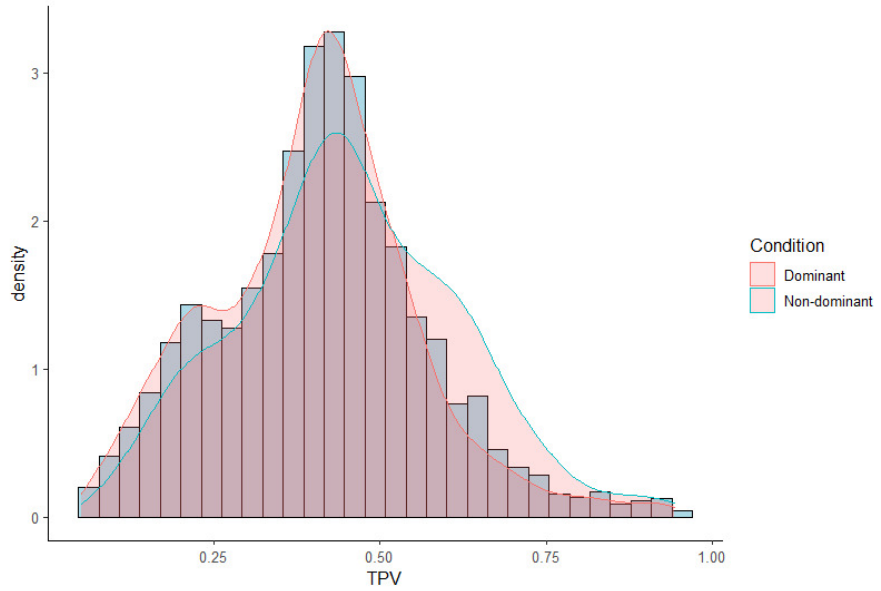


Figure 3.9: Neurotypical adults, distribution of the TPV values ($n_{trials} = 2,903$).

with ADHD provided 2,234 correct and 137 incorrect (i.e., the wrong answer was provided) responses, while children with TD provided 3,777 correct and 211 incorrect responses.

Accuracy Percentages of correct responses according to *Group* and *Condition* are reported in Table 3.6. Model comparison was run with the glmmTMB [147] R package. The binomial distribution was specified to account for the binary nature of the dependent variable (1 = correct, 0 = incorrect). According to Akaike weights ($AICcWt_m0 < .01$; $AICcWt_m1 = .39$; $AICcWt_m2 = .14$; $AICcWt_m3 = .15$; $AICcWt_m4 = .14$), the best model was `m1` (39% probability of being the best model; $\chi^2 = 369.3$; $p < .001$), which revealed a significant effect of *Condition* ($p < .001$). As visualized in Fig. 3.11, accuracy was reduced in the non-dominant condition. Conditional R^2 (the ratio of variance explained by fixed and random effects over total variance) indicates that `m1` explains 33% of variance, whereas marginal R^2 (the ratio of variance explained by fixed effects over total variance) indicates that *Condition* explains 19% of variance. Therefore, 14% of variance was explained by individual variability (i.e., the random effect of participants).

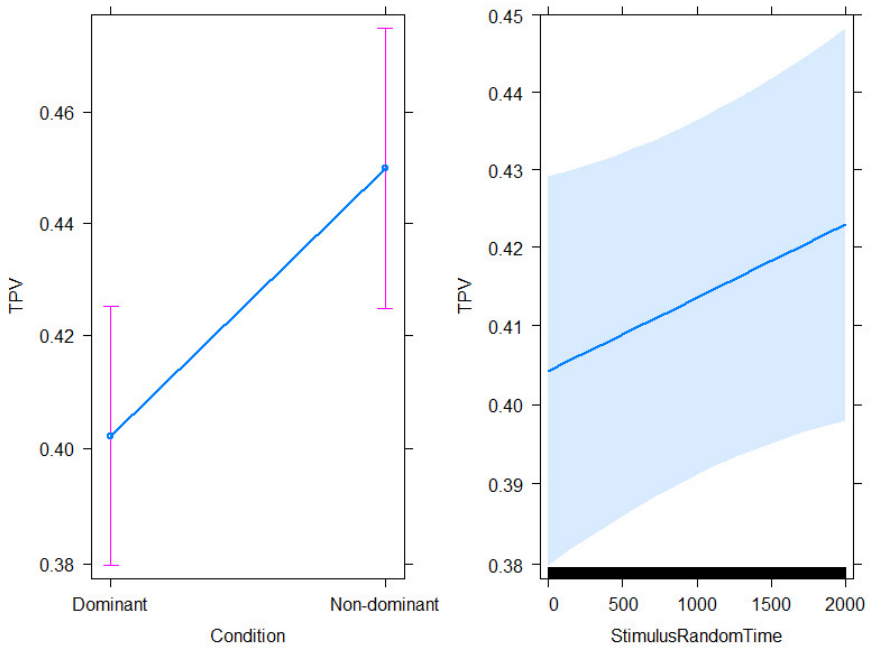


Figure 3.10: Neurotypical adults, model `mb2`, *Condition* and *StimulusRandomTime* effects on the TPV ($n_{participants} = 19$; $n_{trials} = 2,903$).

RT, MD, and TPV We further explored kinematic features of correct responses to investigate whether, beyond accuracy, children with ADHD would show subtle motor atypicalities. Means (M) and Standard Deviations (SD) of RT, MD, and TPV of correct responses in each condition and group are reported in Table 3.7.

In addition, we conducted a visual inspection of the velocity shape and trend across movement time, describing both group and individual differences. At the group level, children with ADHD show a flatter velocity profile over the time course of the movement, with a less evident peak velocity at the beginning of the movement. Due to the high number of images related to this analysis, all the plots and the relative discussion are reported in Appendix 3.A.

RT Model comparison was run with the `glmer` function of lme4 [151] R package. The gamma distribution was specified to account for the positively skewed nature of the dependent variable. According to the Akaike weights ($AICcWt_m0 < .001$;

		Accuracy	
Group	**Condition**	**M**	**SD**
ADHD	Dominant	98%	3%
	Not-dominant	83%	16%
TD	Dominant	98%	2%
	Not-dominant	85%	11%

Table 3.6: Children with ADHD and TD, descriptive statistics of accuracy levels, percentage of correct responses ($n_{ADHD} = 17$; $n_{TD} = 26$).

$AICcWt_m1 < .01$; $AICcWt_m2 < .01$; $AICcWt_m3 = .19$; $AICcWt_m4 = .80$), the best model is `m4` (80% probability of being the best model; $\chi^2 = 4.9$; $p = .03$), which reveals a significant interaction between *Condition* and *Group* ($p = .03$), and a significant effect of *Age* ($p < .001$). As visualized in Fig. 3.12, children with TD showed increased RT in the non-dominant compared to the dominant condition, thus devoting more time to motor planning when the response required inhibition. This pattern was not present in children with ADHD, who did not differentiate RT depending on *Condition*. Moreover, there is a negative association between RT and *Age*, with RT decreasing at older ages, regardless of group. Conditional R^2 (the ratio of variance explained by fixed and random effects over total variance) indicates that `m4` explains 37% of variance, whereas marginal R^2 (the ratio of variance explained by fixed effects over total variance) indicates that *Condition*×*Group* and *Age* explain 28% of variance. Therefore, 9% of variance is explained by individual variability (i.e., the random effect of participants).

MD Model comparison was run with the `glmer` function of lme4 [151] R package. The gamma distribution was specified to account for the positively skewed nature of the dependent variable. According to the Akaike weights ($AICcWt_m0 < .001$; $AICcWt_m1 < .29$; $AICcWt_m2 < .41$; $AICcWt_m3 = .22$; $AICcWt_m4 = .08$), the

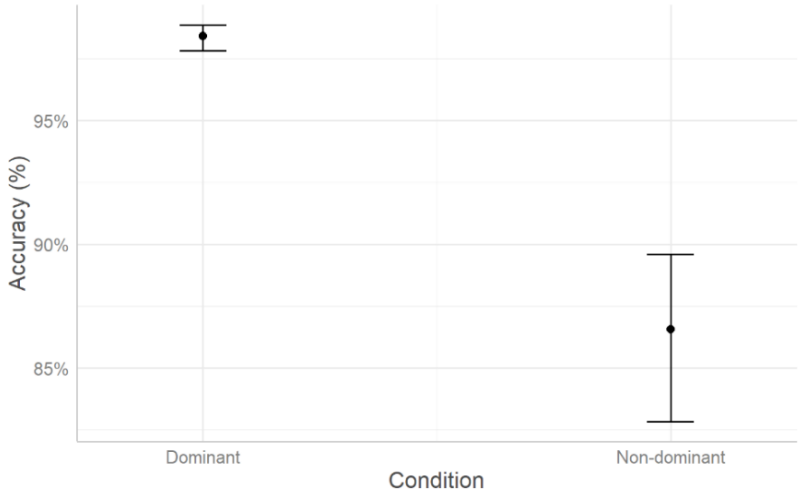


Figure 3.11: Children with ADHD and TD, predicted effect of *Condition* on accuracy ($n_{trials} = 6,359$; $n_{ADHD} = 17$; $n_{TD} = 26$; estimated marginal means with whiskers representing 95% confidence intervals).

best model is `m2` (41% probability of being the best model; $\chi^2 = 2.7$; $p = .1$), which reveals a significant effect of *Condition* ($p < .001$), and a non-significant effect of *Group* ($p = .09$). As visualized in Fig. 3.13, MD increased in the non-dominant condition compared to the dominant condition. Conditional R^2 (the ratio of variance explained by fixed and random effects over total variance) indicates that `m4` explains 38% of variance, whereas marginal R^2 (the ratio of variance explained by fixed effects over total variance) indicates that *Condition* and *Group* explain 20% of variance. Therefore, 18% of variance is explained by individual variability (i.e., the random effect of participants).

TPV Model comparison was run with the glmmTMB [147] R package. The beta distribution was specified to account for the nature of the dependent variable (continuous proportions on the interval 0 : 1). According to the Akaike weights ($AICcWt_m0 < .01$; $AICcWt_m1 = .08$; $AICcWt_m2 = .06$; $AICcWt_m3 = .04$; $AICcWt_m4 = .83$), the best model is `m4` (83% probability of being the best model; $\chi^2 = 8.3$; $p = .004$), which reveals a significant interaction between *Condition* and *Group* ($p = .004$), and a non-significant effect of *Age* ($p = .3$). As visualized in

		RT		MD		TPV	
Group	**Condition**	**M**	**SD**	**M**	**SD**	**M**	**SD**
ADHD	Dominant	652	211	565	217	447	182
	Not-dominant	653	217	734	242	456	228
TD	Dominant	691	198	514	190	460	175
	Not-dominant	716	217	656	242	504	215

Table 3.7: Children with ADHD and TD, descriptive statistics of correct responses, values in ms ($n_{trials} = 6,011$; $n_{ADHD} = 17$; $n_{TD} = 26$).

Fig. 3.14, TD children showed increased TPV in the non-dominant compared to the dominant condition, thus devoting more time to motor planning when the response required inhibition. This pattern was not present in children with ADHD, who did not differentiate TPV depending on *Condition*. At both group and individual level, further graphical inspection of velocity shape across time is described in Appendix 3.A. Conditional R^2 (the ratio of variance explained by fixed and random effects over total variance) indicates that `m4` explains 71% of variance, whereas marginal R^2 (the ratio of variance explained by fixed effects over total variance) indicates that *Condition*×*Group* and *Age* explain 9% of variance. Therefore, 62% of variance is explained by individual variability (i.e., the random effect of participants).

3.8 Discussion

3.8.1 Neurotypical adults

In the first study, we explored neurotypical adults' movements during the implemented task, an adapted *Go/No-Go* paradigm (details in Section 3.4.4) [3].

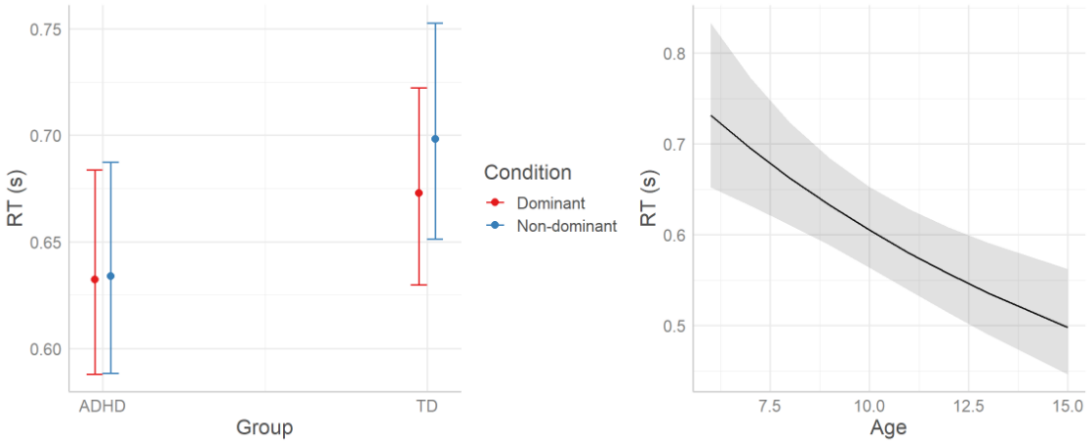


Figure 3.12: Children with ADHD and TD, predicted effects of *Condition*×*Group* and *Age* on RT ($n_{trials} = 6,011$; $n_{ADHD} = 17$; $n_{TD} = 26$; RT is expressed in seconds; estimated marginal means with whiskers representing 95% confidence interval; for the *Age* effect shaded area represents the 95% confidence interval).

Motor planning and motor control The descriptive statistics indicated that participants performed the non-dominant response (compared to the dominant one) by increasing both the RT (Reaction Time, time devoted to motor planning) and MD (Movement Duration, time of motor execution). However, these two indices are not sufficient to disentangle the planning and control phases of the movement. Indeed, given that motor planning and control overlap during the MD [94], we analyzed the Time to Peak Velocity (TPV) to further distinguish these two mechanisms. As a relative asymmetry index, whether the TPV occurred earlier or later over the MD would reflect more either planning or control. From our exploratory model comparison, we can expect people to show bigger TPV in the non-dominant compared to the dominant condition. This evidence supported the idea that neurotypical adults require greater motor planning rather than an online adjustment to inhibit a prepotent response, selecting and performing an alternative one. Our results are consistent with the extant literature, whereby planning is devoted to processing cognitive information and control is dedicated to getting on a target and adjusting to its specific spatial features [94].

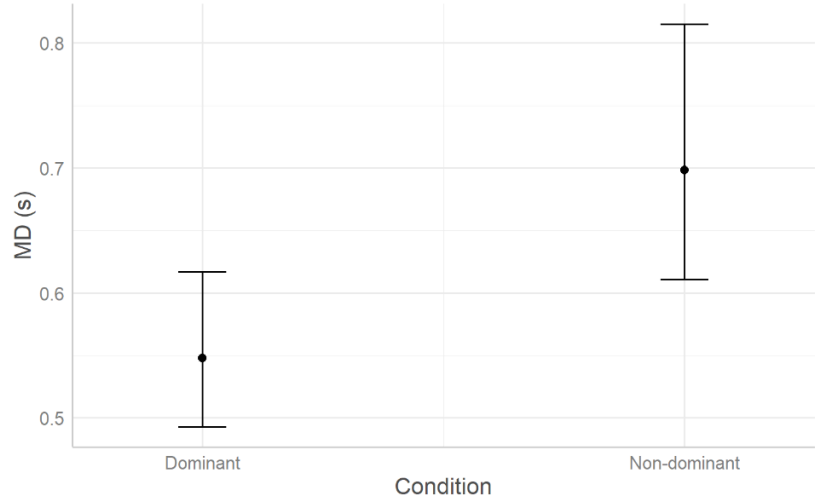


Figure 3.13: Children with ADHD and TD, predicted effects of *Condition* on MD ($n_{trials} = 6,011$; $n_{ADHD} = 17$; $n_{TD} = 26$; MD is expressed in seconds; estimated marginal means with whiskers representing 95% confidence interval).

Stimulus random time The most plausible model given our data and set of specified models showed that when people had to wait for more to start the trial (*StimulusRandomTime*), they increased the movement time devoted to motor planning. Although not significant from a statistical point of view, this effect suggests that a longer preparation time before the trial starts might allow participants to increase the time devoted to motor planning. We can interpret this finding in light of the massive literature about the preparatory effect of the fore period, which is the time from a warning signal and a "Go" stimulus, and is known to affect response times [152]. In our study, participants had to place their hand on the presence sensor to signal their readiness to start the next trial. The time instant they pressed the sensor can be seen as an active warning signal that pre-activates the sensorimotor system. After a variable random time interval (*StimulusRandomTime*), the central "Go" stimulus appeared to trigger participants' responses. We can speculate that, within 2,000 ms, a longer preparation time increases adults' motor planning. As the fore period effect and the temporal preparation abilities change across development, future studies could expand on the ontogeny of these mechanisms [153].

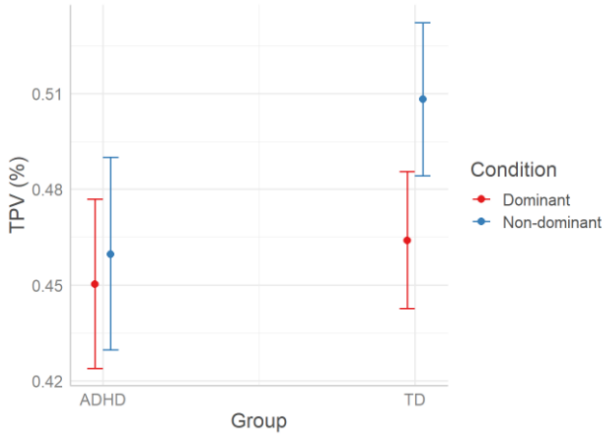


Figure 3.14: Children with ADHD and TD, predicted effects of *Condition*×*Group* and *Age* on TPV ($n_{trials} = 6,011$; $n_{ADHD} = 17$; $n_{TD} = 26$; TPV is expressed as a percentage within the 0 : 1 range; estimated marginal means with whiskers representing 95% confidence interval).

Limitation It is worth mentioning that this first study has some limitations: (a) our sample did not include a balanced number of women and men, thus preventing us to make any claims about potential gender differences, that should be furthered in future studies, and (b) we could not base our sample size specification on previous literature that tested motor inhibition through the TPV%. Therefore, our findings should be interpreted as preliminary and exploratory indications to develop future confirmatory studies.

3.8.2 Children with TD and ADHD

In a second study, we explored the mechanisms underlying the inhibition of a prepotent motor response, which is frequently reported to be affected in children with Attention Deficit and Hyperactivity Disorder (ADHD) [6]. The performance of ADHD and Typical Development (TD) groups at our adaptation of a *Go/No-Go* paradigm (details in Section 3.4.4), showed both similarities and differences.

Accuracy Both children with ADHD and TD made more errors in the non-dominant compared to the dominant condition. This indicates that the task was effective in inducing a prepotent response in the dominant condition, which was the more frequent one, and facilitated by the requirement to match the "Go" stimulus and the response option by color. Children with ADHD and TD were equally accurate in selecting the correct response so no group difference was found in accuracy levels. This unexpected result could be due to the ease of the task, which required a rather simple motor response, as also evidenced by the high percentages of correct responses. In tasks with greater time pressure or greater complexity of the motor action required to answer, we could expect more marked differences between the two groups. Although the task was based on the central properties of *Go/No-Go* (i.e., more frequent administration of the dominant condition), some differences may have made our task easier than traditional ones at the level of inhibition of prepotent responses. In particular, responding by reaching rather than quickly pressing a button may have allowed participants more time to process the cue, recall the instructions, and redirect their response during movement. In addition, this may explain the high accuracy, and at the same time allowed us to study not only the reaction time (movement pre-planning) but also what happens during movement (motor planning gradually gives way to control of the ongoing movement).

Motor planning Beyond accuracy, the main findings of this study revealed that the ADHD group showed different motor patterns that possibly indicate reduced motor planning compared to the TD group. In the non-dominant condition compared to the dominant condition, TD children spent more time planning the movement, which resulted in longer Reaction Time (RT) and greater percent Time to Peak Velocity (TPV%). Indeed, a higher relative time to peak velocity, i.e., greater TPV, is an efficient strategy of the motor system, that reduces the time and resources needed for online movement correction [154]. In addition, children with ADHD did not modulate RT and TPV according to condition, not dedicating more time to motor planning when needed to inhibit the prepotent response.

This subtle lack of flexibility in adjusting the motor and cognitive strategies to the task demands can be interpreted as a marker of motor and cognitive impulsivity. Our findings are in line with previous literature showing that atypical activation of premotor systems may contribute to impaired response inhibition in children with ADHD [155]. There is an interesting debate in the literature on the link between motor preparation and spatial attention [156, 157], which could be further explored to understand the link between cognition and movement in ADHD.

Motor control Across both groups, children showed increased Movement Duration (MD) in the non-dominant vs. dominant condition. This indicates that inhibitory processes take place during movement execution. Throughout the movement, motor planning gradually gives way to control and monitoring of the ongoing movement [94]. To better disentangle how much of the movement time is devoted to planning or control, we employed the percent Time to Peak Velocity (TPV%) as a relative asymmetry index. Theoretical reaching trajectories starting and ending at full rest have a bell-shaped velocity path, with the first half of MD spent accelerating and the second one decelerating, resulting in a 50% TPV [122, 121]. In actual reaching movements, distinct characteristics of the target differently affect the movement acceleration-deceleration symmetry. Whereas physical precision of the movement (e.g., grasping small objects) requires more control and longer deceleration [158], cognitive load affects the early stages of movements, thus requiring more planning [159]. The smaller TPV captured across conditions in the ADHD compared to the TD group indicates a higher portion of movement being dedicated to the deceleration phase, which usually stands for motor control [94, 158]. Increased movement variability in children with ADHD [160] has often been interpreted as an indication of poor motor control, when instead it could be a compensatory strategy that, given a reduced planning, requires more online adjustments during movement execution. To better understand "how" children with ADHD regulate movement in its final phase, future studies would benefit from the use of additional kinematic indices that capture online motor correction more precisely (e.g., the number of direction changes and acceleration/deceleration units).

Although kinematic indices are widely used as a mirror of underlying cognitive mechanisms, it is even more informative to combine them with the study of neural correlates [161]. Previous evidence suggests that increased activation of prefrontal areas can help children with ADHD compensate for atypical activation of premotor areas in *Go/No-Go* tasks [155]. It is unclear whether this can be attributed to planning or control mechanisms. The study of EEG components and the timing of neural activities that precede and take place during responses to cognitive tasks can be coupled with kinematic indices to shed light on planning-control dynamics [162, 163]. To the best of our knowledge, little is still known about the specifics of such mechanisms in ADHD.

Children with ADHD might employ compensatory strategies for planning difficulties, which may be sufficient to achieve good accuracy in very simple tasks as the one employed in our work. Indeed, they chose between two alternatives that differed only in one motor (i.e., the movement direction: reaching the key to the right or to the left of the central stimulus) and cognitive (i.e., the response key color) parameter. However, this might not be sufficient in more naturalistic situations, in which alternative choices differ in more complex kinematic parameters (e.g., using the right arm or the left arm to respond), or require finer cognitive processing (e.g., selecting the most appropriate behavior according to a specific social context).

Age From the obtained results, it is also possible to see a progressive reduction in RT as the age of participants increases, which is consistent with decades of findings from developmental studies [164]. This suggests that motor planning becomes globally more effective and rapid with age, and therefore requires fewer cognitive resources. Given the low sample size, the statistical models tested included the age variable as an additive effect (i.e., irrespective of experimental condition and group membership). Thus, we accounted for the differences attributable to the age of participants in the accuracy and overall kinematic profile. However, we did not specifically assess the role of age in interaction with the other predictors (i.e., experimental condition and group membership).

Limitation It is worth mentioning that also this second study has some limitations. As we were not interested in assessing gender differences, (a) our sample is not balanced by participant gender, which reduces its representativeness of the general population. In addition, the sample size was determined by the number of families that agreed to participate in the study. Given the complexity of the experimental design (i.e., multiple dependent and independent measures are of interest), its exploratory nature, and the paucity of prior evidence on which to estimate expected effect sizes and appropriate sample sizes, (b) our sample size may be insufficient to reveal further differences between groups. Further inferential research will be needed to confirm the considerations presented in this work. Nevertheless, research on developmental populations with specific conditions frequently suffers from small sample sizes and even single-case studies. Replication of studies, meta-analyses, and multi-lab projects would help deal with this issue in the long run of knowledge acquisition, whereby every study contributes to a piece of the puzzle.

3.9 Conclusions

The present work wants to investigate the relative contribution of motor planning and control to the inhibition of a prepotent response. Before starting with the analyses, we concentrated on implementing a task to assess the selection or inhibition of a prepotent response, and a *Go/No-Go* paradigm was adapted to this. Successively, we focused on the experimental procedure, task and apparatus, and methods to obtain reliable results. Importantly, in every single phase of the research, we always proceeded with the aim of setting a stage for a machine learning task, waiting to reach a sufficient amount of data.

In addition, the present work employed a low-cost wearable 3-axis accelerometer to investigate human motor inhibition. The inertial sensors built with 3-axis accelerometers, gyroscopes, and magnetometers have been indicated as promising commercial tools to study the kinematics of human movements and overcome the constraints of expensive motion capture systems. Although they have the poten-

tial of being portable and wearable, they appeared to provide accurate and reliable data only for some kinematic indices, such as the value and timing of peak velocity [112]. On one hand, based on our kinematic measurements and analyses, the kinematic indices built upon the velocity value did not appear sufficiently reliable and valid (Section 3.6.2). On the other hand, however, those related to the velocity shape over time seemed to be valid indices. Indeed, our average Time to Peak Velocity percentage (TPV%) was consistent with those reported by previous studies, similar tasks, and motion capture systems with the highest level of precision [143]. *Therefore, we support the use of a commercial and low-cost* 3*-axis accelerometer to calculate the TPV% and compare participants' performance.* Then, future studies could utilize the present method and apparatus to disentangle the planning and control mechanisms of motor actions that involve different neuropsychological abilities, thus providing fundamental insights into the design of motor and psychological interventions.

3.9.1 Neurotypical adults

Overall, the first study of this work expands on our understanding of which motor strategy is successful for neurotypical adults to inhibit prepotent reaching movements [3]. This would lay the foundations for investigating the atypical strategies implemented by individuals and clinical groups with inefficient motor inhibition. Although motor inhibition is affected in several neurodevelopmental disorders, the underlying multifaceted mechanisms shape unique phenotypes that require appropriate and specific interventions [165]. For instance, inhibitory skills are linked to individual traits such as impulsiveness [166], and inhibitory control deficits have been found through *Go/No-Go* tasks in autism spectrum disorders [167], whereby difficulties in inhibiting prepotent responses seem to be associated with higher-order repetitive behaviors [168]. Moreover, inhibition is part of a broader category of control processes named executive functions, which are distinguished but correlated [169], and play a fundamental role in everyday action selection and execution. Indeed, although difficulties and impairments in the action domain are common to several clinical conditions (i.e., multiple sclerosis,

Alzheimer's disease, Parkinson's disease), the underlying sensory, motor and cognitive mechanisms might dramatically differ among patients [170, 171, 172, 173].

3.9.2 Children with TD and ADHD

In a second study, instead, we explored the motor strategy to inhibit prepotent reaching movements for children with ADHD and TD, considering the same experimental procedure, task, and apparatus [4, 5, 6]. Children with ADHD can exhibit similar accuracy to the TD control group in simple tasks tapping on the inhibition of prepotent motor responses. However, accurate inhibition appears to be achieved through different mechanisms, including less motor planning and greater ongoing control of movements. Although online control of one's own responses may be sufficient to compensate for planning difficulties in simple experimental tasks, this could profoundly impact the behavior of children with ADHD in everyday life contexts, which involve very complex choices among numerous possible alternatives. Moreover, motor and cognitive impulsivity might be related to broader atypicalities, ranging from sensory atypia and stereotypies to executive difficulties in everyday tasks. For this reason, it is fundamental to understand the mechanisms underlying impulsivity and design interventions that are individualized on the child's profile, and synergistically target the motor and cognitive dimensions of inhibition. To this end, the use of portable, user-friendly, and low-cost kinematic sensors (e.g., a wrist-worn accelerometer) offers great possibilities for neuropsychological assessment and treatment, being also affordable for local clinical services. In sum, this study opens the door to further research that will help the scientific and clinical community understand and target impulsivity, leading to benefits on children's developmental trajectory and well-being.

3.10 Future works

Starting from this work, further research is needed to investigate the implications of atypical motor and cognitive inhibition on the daily life, learning, and social skills of children with ADHD. Future studies with appropriate sample sizes and

broader age ranges may further investigate developmental changes in inhibitory strategies, also exploring potential ADHD-related differences. For instance, some children with ADHD show stereotypies which are involuntary, restricted, and repetitive patterns of behaviors that limit the child's resources to learn and practice various, appropriate and goal-directed actions [174, 175, 176]. Specifically, motor stereotypies are present in both neurodevelopmental conditions and typical development [177], and might be related to ineffective motor planning [178] and inhibitory difficulties [179]. Indeed, motor-related cortical potentials in premotor areas, which anticipate voluntary motor actions, are found to be absent before stereotypy onset in typical development [178]. Stereotypies are mostly studied in Autism Spectrum Disorder (ASD), as they are core symptoms of those conditions [105]. However, they are frequently found in ADHD, and show similar characteristics across ASD and ADHD [180], which often co-occur, share clinical manifestations, and entail impairments in overlapping mechanisms [181, 182]. Notably, stereotypies can be related to cross-diagnostic sensory, motor, and cognitive mechanisms. Atypical inhibition of prepotent responses is correlated with repetitive behaviors, with differences between higher-order and sensorimotor stereotypies [168, 183]. Moreover, stereotypies are associated with sensory difficulties [184], which can be present in ADHD [185, 186, 187], and are bounded to the motor and cognitive processes through complex, dynamic, and multidirectional relationships. We can speculate that those children with greater stereotypies could have less effective sensory and executive profiles, as well as motor planning difficulties. They might need to devote more resources to motor control to effectively inhibit a prepotent response. In [188], for example, our works [3, 6] are cited for research regarding agency and reward across development and in autism. The authors, in particular, aim at disentangling the role of agency and reward in driving action selection of autistic and non-autistic children and adults, considering different variables (among which reaction time measures) and a free-choice paradigm. Future studies may employ our paradigm to better understand whether atypical cognitive and motor inhibition may contribute to broader individual differences in everyday sensory, cognitive, and social functioning. Studies with more hypothesis-driven

approaches and an appropriate sample size would allow us to draw clearer, more inferential conclusions on the complex relationships between these variables.

Shifting to the experimental apparatus considered, this work opens the door to important application challenges in bringing these methods and knowledge into clinical practice. It would be crucial to integrate the kinematic analysis into the classical neuropsychological tests that evaluate executive functions, to better understand how a response to a given test is planned and adjusted along the way. In this regard, the distinction between reaction time and movement duration is a promising perspective for neuropsychological research, as it allows a distinction to be made between two different mechanisms underlying a response (i.e., planning and control). Moreover, this method would facilitate not only the identification of specific difficulties and the monitoring of the treatment effects but also serve as an intervention tool itself. For instance, using kinematic measures as biofeedback could promote patients' awareness of their behaviors and facilitate learning strategies to modify them. Although the use of inexpensive and portable kinematic sensors removes one of the barriers to its use in the clinic, the difficulty of analyzing and interpreting the raw data obtained with such instruments remains. To overcome this obstacle, it will be necessary for researchers to develop and make available user-friendly software that processes the raw kinematic data and compute performance indices that are interpretable by clinicians. To this end, we first need large-scale validation studies that provide normative values and risk indices to evaluate an individual's performance.

Along the same line, from a performance perspective, future studies would benefit from the use of additional kinematic indices that capture online motor correction more precisely, e.g., number of direction changes, acceleration/deceleration units. Again, future studies might include video recordings and offline coding of the experimental sessions, thus (a) checking for potential cases where participants show extra-task movements that could result in anomalous trials, and (b) further increasing the accuracy of the preprocessing from a methodological point of view. In particular, to remove the gravity component from the acquired acceleration, future studies could use a combination of an accelerometer and gyroscope. In this

way, data related to the orientation of the accelerometer would be available in order to remove the gravitational acceleration. However, the gyroscope would not solve the numerical errors driven by possible accelerometer bias and numerical mathematical functions. These issues could be addressed from an algorithmic point of view, with the evaluation of other methods and models in order to process raw accelerometer data in a way that could reduce the numerical errors. Here is the key point: the learning algorithms, an algorithm class that could obtain promising results with huge amounts of raw data.

Machine learning algorithms could study different input signals and learn information from the data. In this case, a supervised data set would incrementally improve the results, but also an unsupervised approach could be taken into account. The idea of applying machine learning algorithms is where "we arrive" from this work, but it is also from where "we start". Importantly, in fact, not to forget the idea of this work of setting from scratch the stage for a machine learning task to explore a specific aspect of human reaching movement (the distinctive contribution of motor planning and control) through an accelerometer-based analysis.

3.A Appendix - Velocity shape and trend

We here conduct a visual inspection of our data, with a specific focus on velocity shape and trend across movement time [6]. Notably, only correct trials (i.e., trials in which the participant gave the correct answer) are considered. Firstly, we plotted the data of the two groups separately, children with Attention Deficit and Hyperactivity Disorder (ADHD) and with Typical Development (TD). Secondly, we plotted individual data to explore individual variability. As explained in the previous Section 3.6.2, we do not focus on velocity magnitude values, but rather focus on velocity curve shape and trend in time [3].

Group level At the group level, i.e. children with ADHD and TD, each Figure is composed of 3 graphs, one for each row. The first and second graphs constitute a boxplot composition from trials in either the dominant (red) or non-dominant (blue) condition, respectively. The x-axis represents the movement time (in ms), whereby the instant in which the participant starts moving (R) is aligned with the 0 value. The y-axis, instead, shows the velocity values (in m/s). For each 10 ms of movement time (corresponding to the accelerometer sampling rate), we plotted a boxplot composed of data from all equivalent time points of the different trials. Although the y-axis value ranges were affected by outliers, they were excluded from the visualization for the sake of graphic clarity and readability. For instance, in case some blank spaces appear in the superior and inferior parts of a boxplot, some invisible outliers are present. As we focus on the velocity shape and trend across time, and do not aim to compare its magnitude across different graphs, we did not set a fixed y-axis range for all the Figures. We have therefore avoided a flattening of the boxplots resulting from variability between participants. As not all the trials have the same movement duration, the boxplots are composed of varying amounts of data. We consider this fact in the third graph, which represents the number of trials contributing to each time instant of each boxplot, in either the dominant (red) or non-dominant (blue) condition, respectively.

Individual level At the individual level, in addition to the graphs reported at the group level, we reported a plot representing velocity for each trial. Each curve corresponds to a single trial and is visualized in red for the dominant condition, and in blue for the non-dominant condition. The x-axis represents the movement time (in ms), while the y-axis shows the velocity values (in m/s). Then, a vertical green line marks the time instants when the participant ended his/her movement in each trial (A), the participant touches the screen and provides an answer. Green lines make it easier to capture movement duration in each trial.

3.A.1 Children with TD and ADHD - Group analysis

From a group-level visualization, it is possible to observe that the TD group showed a more pronounced bell-shaped velocity pattern at the beginning of movements, as represented by the time-ordered set of boxplots (Fig. 3.15). This seems in line with previous literature suggesting that children with ADHD do not show a typical bell-shaped velocity profile, which indicates impaired motor planning [128].

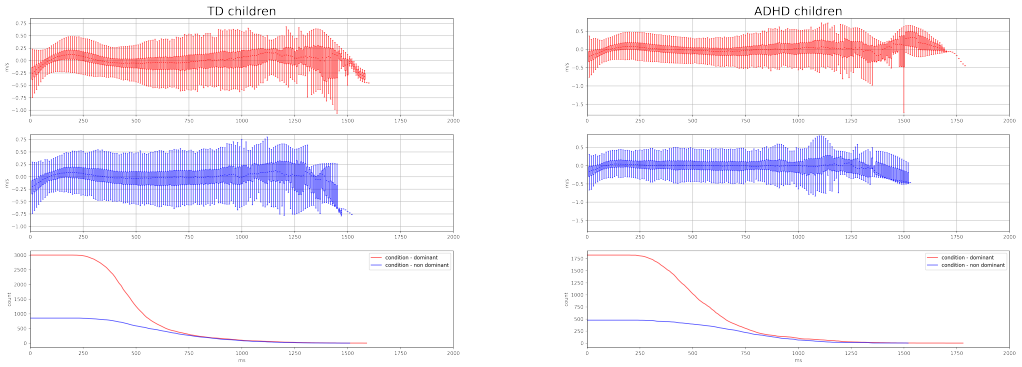


Figure 3.15: Children with TD and ADHD groups.

3.A.2 Children with TD - Individual analysis

For the 26 children with TD participants, numbers ranging from 101 to 126 are reported as a participant's identification code (Figs. 3.16, 3.17, 3.18, 3.19, 3.20). From these graphs, it is possible to note that, although all participants showed a bell-shaped velocity profile at the beginning of movements, there is wide intra-group variability. The significance of this motor variability, concerning the individual characteristics of typically developing children, is largely under-studied in the scientific literature and deserves further investigation to explore the possibility of capturing predictive cues about children's motor development.

3.A.3 Children with ADHD - Individual analysis

For the 17 children with ADHD participants, numbers ranging from 1 to 17 are reported as a participant's identification code (Figs. 3.21, 3.22, 3.23). From these graphs, it is possible to observe that some participants of this group did not show an initial bell-shaped velocity pattern (see participants 1, 2, 3, 11, 14). As for the TD group, profound intra-group variability is visible and would be worth further investigation. We can speculate that, beyond diagnosis, individual differences in children's motor developmental trajectory interact with other neuropsychological domains to delineate risk profiles that merit clinical attention.

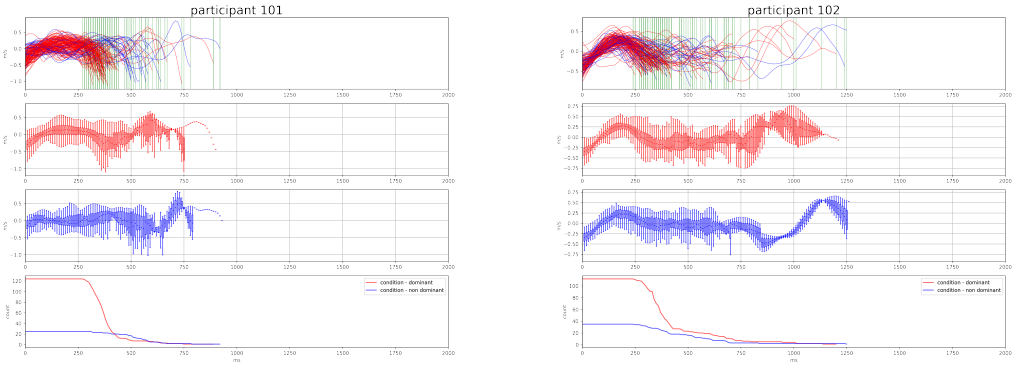


Figure 3.16: Children with TD (part 1).

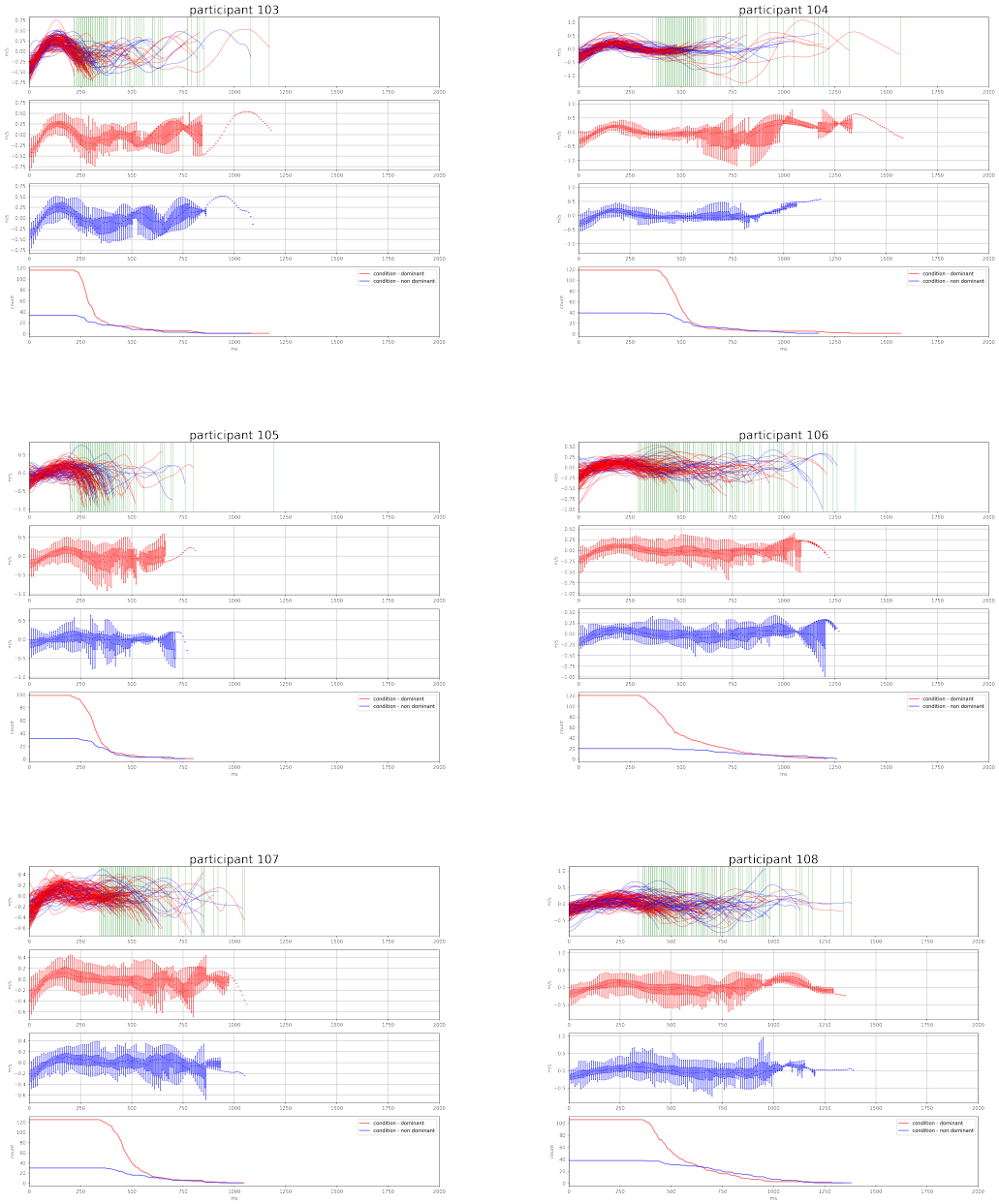


Figure 3.17: Children with TD (part 2).

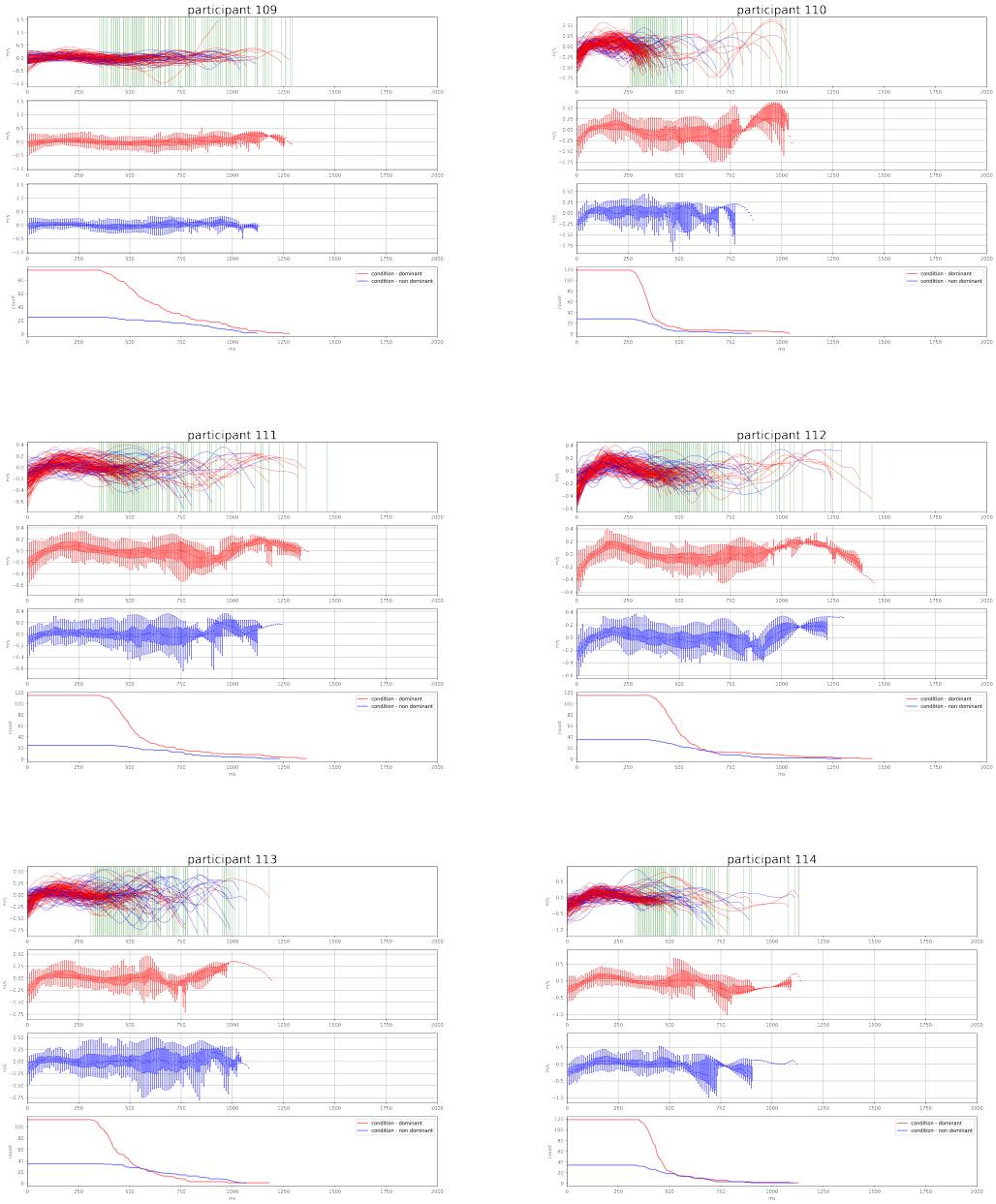


Figure 3.18: Children with TD (part 3).

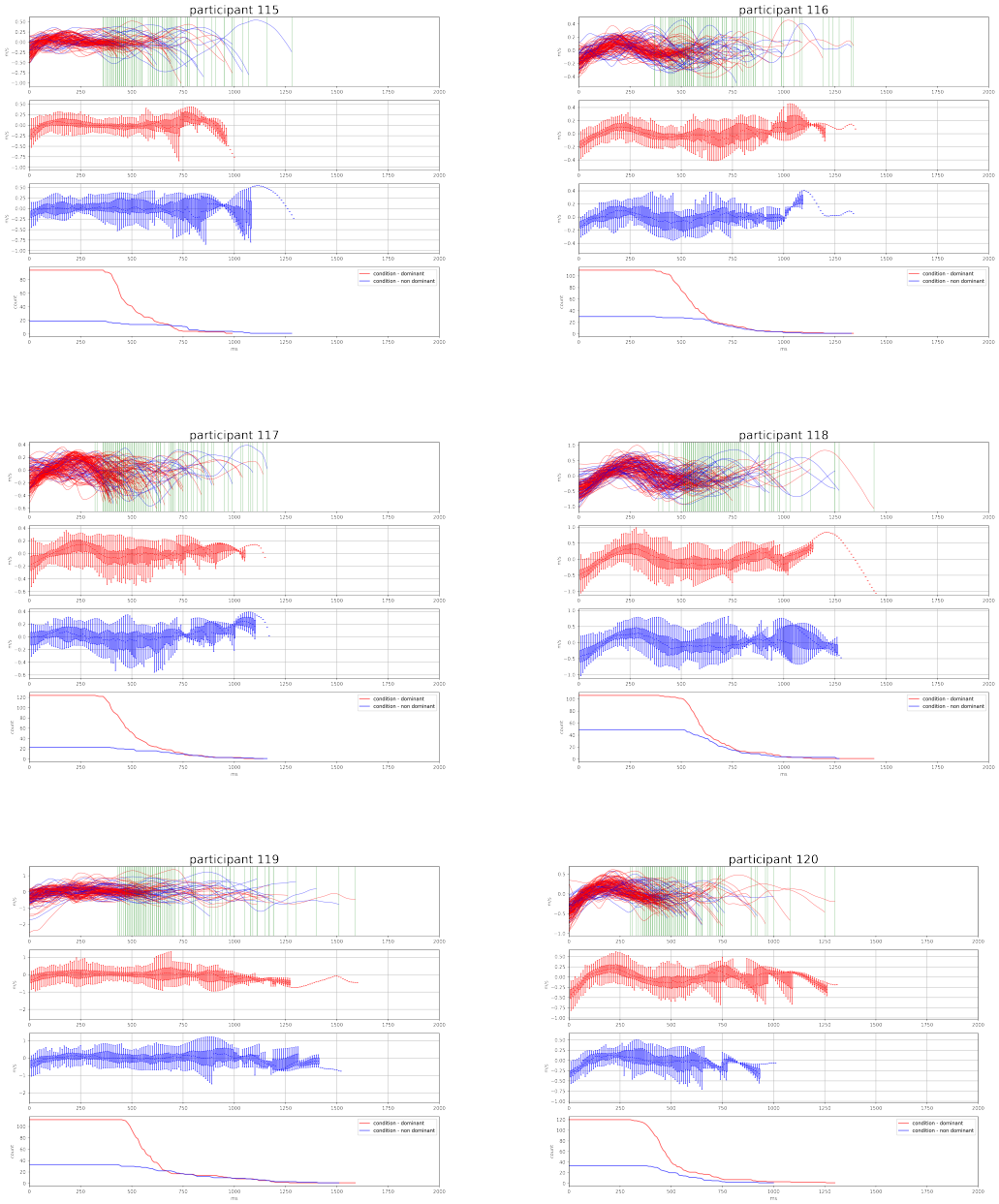


Figure 3.19: Children with TD (part 4).

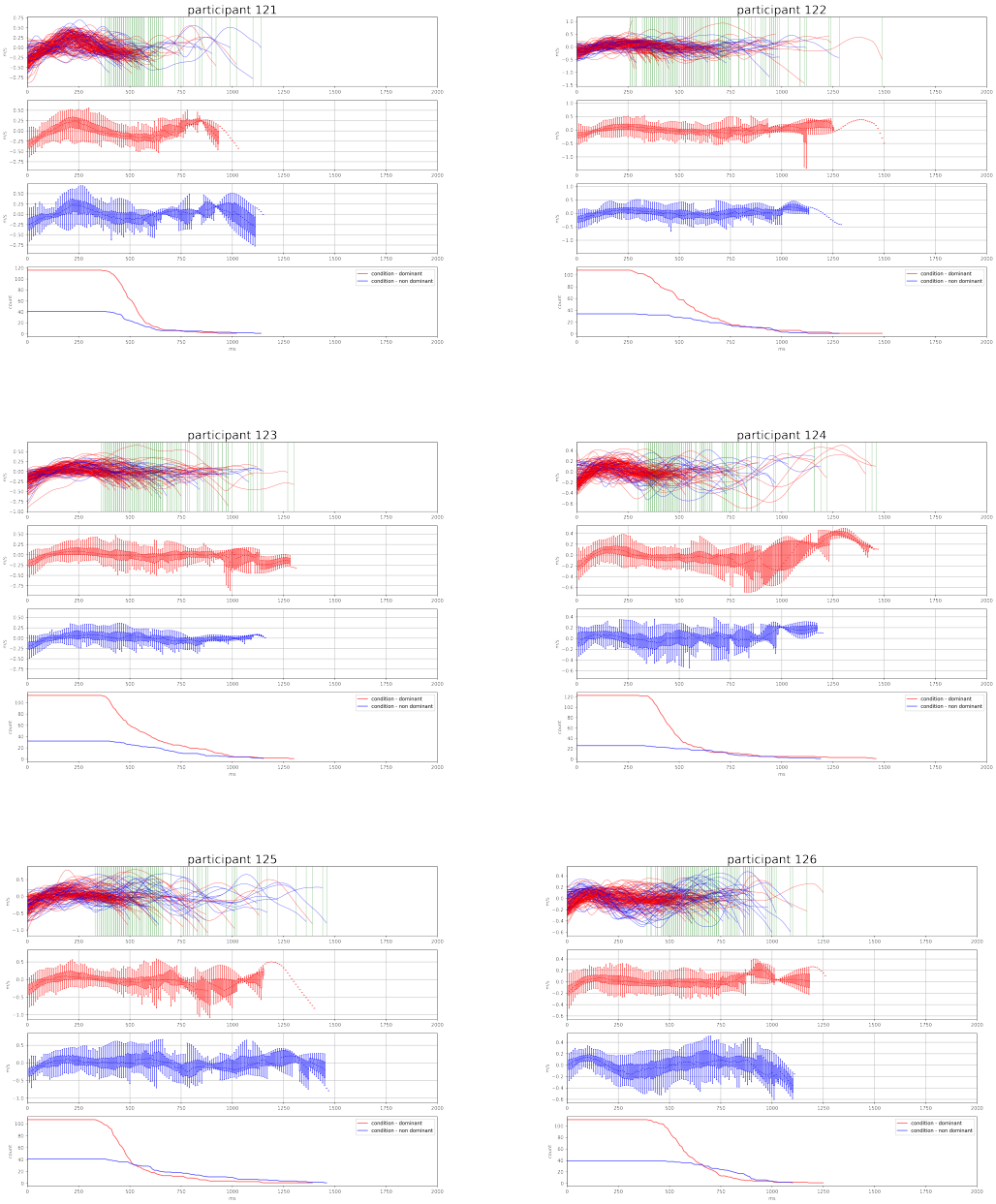


Figure 3.20: Children with TD (part 5).

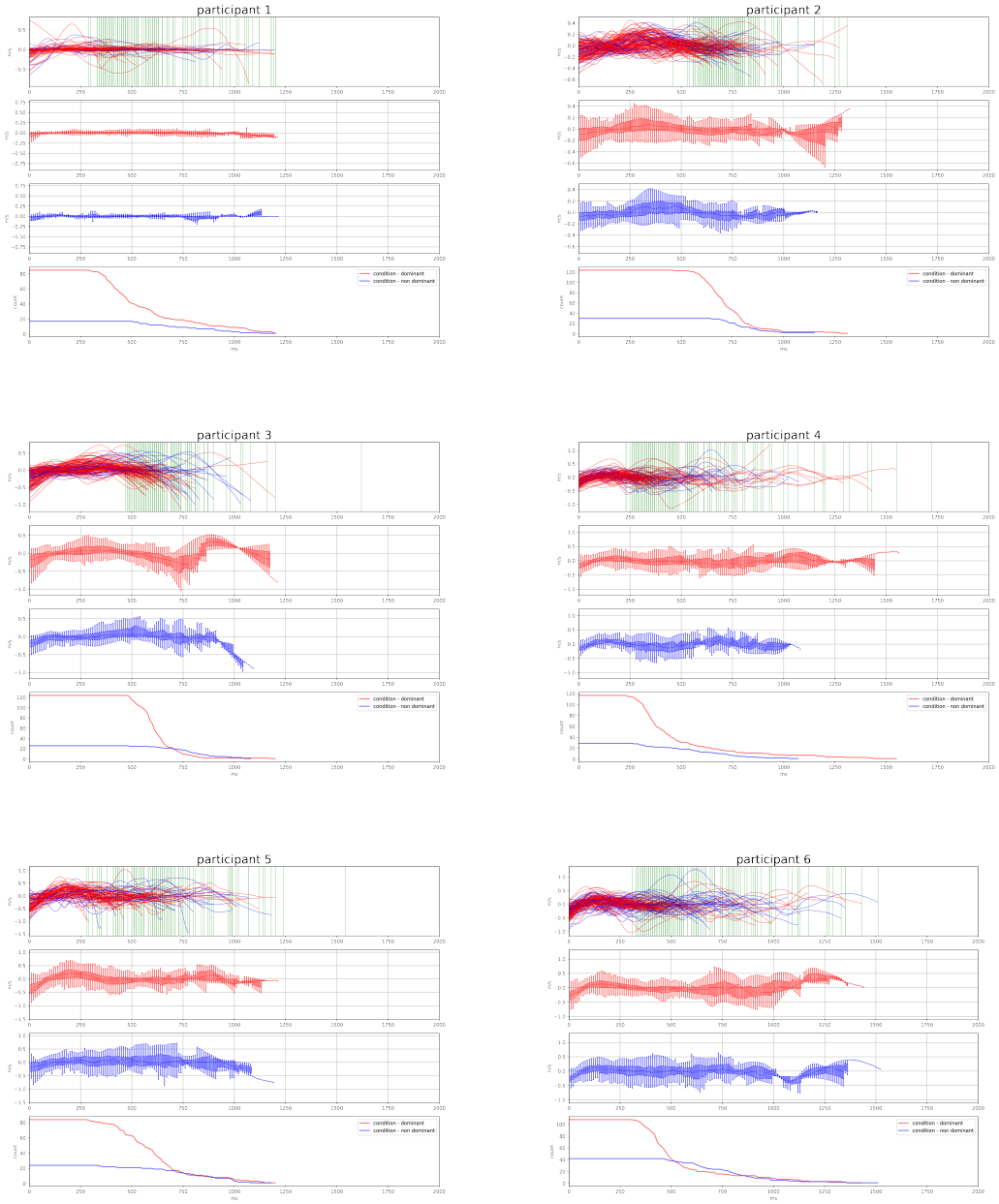


Figure 3.21: Children with ADHD (part 1).

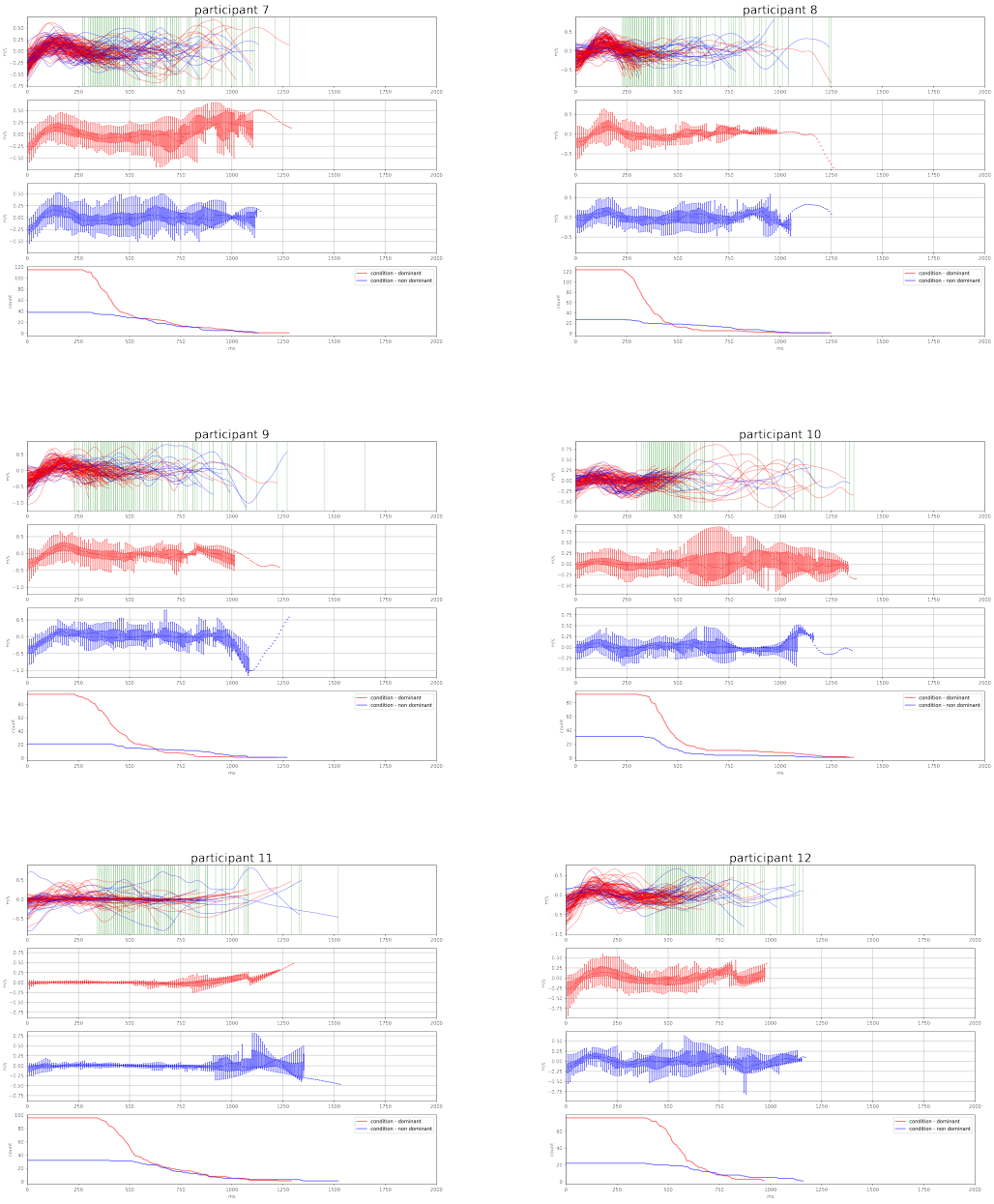


Figure 3.22: Children with ADHD (part 2).

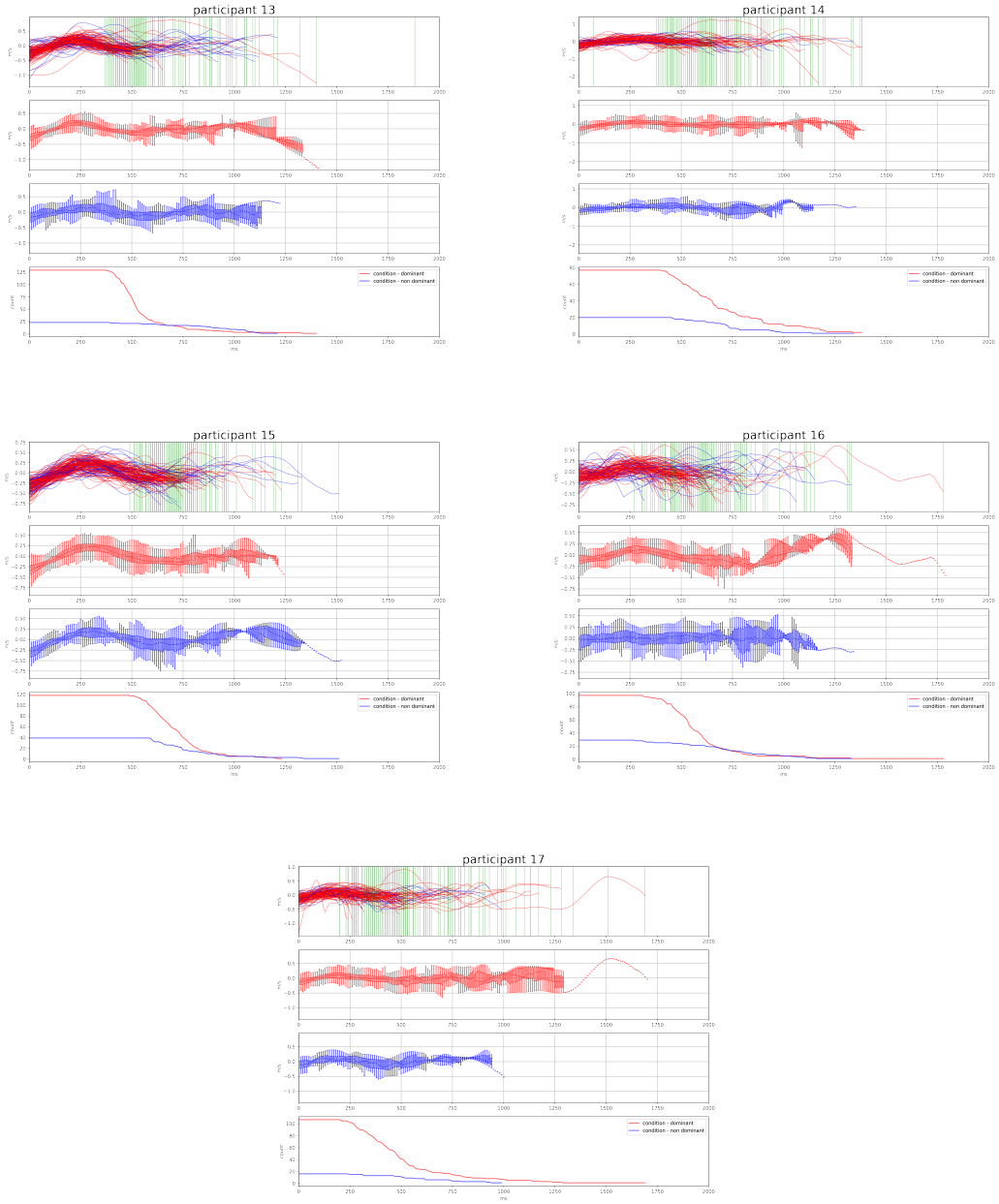


Figure 3.23: Children with ADHD (part 3).

Chapter 4

A detailed machine learning application

Considering Socio-Historical aspects in the Cultural Heritage environment

4.1 Introduction

Analog or digital photography? All started with analog photography, but following Kodak's invention of the first megapixel sensor in 1986, digital photography has slowly grown to substitute its analog predecessor, playing a key role in the early 21st century digital revolution and social transformation [189, 190]. As a relevant example, photography has modified the way mobile phones are used, as their integration of digital cameras has at once fostered an exponential growth of the photos that are shot and uploaded to the Internet every year, as well as a paradigm shift in mobile communications, which today rely on high-quality multimedia [191, 192, 193, 194]. These phenomena have proven to be game-changers for how people communicate and the bloom of new fields of research, as both academia and industry, have exploited such plethora of visual data to develop and apply computer vision models to a variety of different problems (e.g., face recognition, autonomous driving) [195, 196, 197, 198, 199, 200]. Now, while a wealth of research is being devoted to the processing and analysis of digital images, much

has to be done regarding analog ones, mainly because printed images may be: (a) scattered in numerous public and private collections, (b) of variable quality, and (c) damaged due to hard or continued use or exposure. In addition, any analysis utilizing image processing and computer vision algorithms require the potentially quality degrading initial digitization step. Nevertheless, despite the complications and challenges brought on by analog photographs, they still represent an unparalleled source of information regarding the recent past. In fact, no other visual media has been used as pervasively to capture the world throughout the 20th century, as the availability of consumer grade photo cameras supported the spread and popularity of vernacular photography practices (e.g., travel photos, family snapshots, photos of friends and classes) [201, 202].

Family photo albums represent an example of vernacular photography that has drawn the attention of researchers and public institutions. Although one of the most popular practices in photography since the end of the 19th century, an increase in scholarly interest in family photo albums dates back to the early 1980s. A recent work defines family photo albums *a globally circulating form that not only takes locally specific forms but also "produces localities" that create and negotiate individual stories* [203]. Along the same lines, in another relevant contribution, family photo albums *represent a reference point for the conservation, transmission, and development of a community Social Heritage* [204]. In essence, scholars from different fields agree in identifying such type of photography collections as capable to reveal sociological and historical insight and capturing salient features regarding the evolution of local communities in space and time. They are, however, in most cases scattered among private homes and only available on paper or photographic film, thus making their collection and analysis by historians, socio-cultural anthropologists, and cultural theorists very cumbersome. Their study may also become difficult due to the number of photos that such collections contain: a large-scale analysis of such photos, in fact, is often impossible, as manual verification of the characteristics of more than a few hundred pictures would be excessively burdensome, considering also that in many cases no associated descriptions are available. This is why contributions in this field often base their findings on the study of

small corpora of photos [203, 204]. Though, computer-based methodologies could aid such a process in various ways (e.g., speeding the cataloging step) with the use of modern computer vision techniques.

In this work, we investigate such an approach, taking as a case study the socio-historical analysis of a collection of family album photographs. More in detail, we here start introducing the IMAGO dataset, a collection of family album photos assembled and maintained at the University of Bologna, since 2004 [204]. Then, we proceed to present the design and implementation of a multimedia application that, resorting to deep learning models, implements the family album photo classification for cataloging purposes. Exploiting the proposed application, the IMAGO dataset has offered the opportunity of experimenting with photos taken between the years 1845 and 2009, characterized by the fact that each image portrays at least one person. In particular, it has been possible to estimate their socio-historical content, i.e., the shooting dates and socio-historical contexts of the images, through a deep learning-based approach without resorting to any other sources of information.

Nevertheless, this is not all. In fact, starting from this project we proceeded with additional analysis and applications. More in detail, (a) focusing the analysis from a different perspective, we verified whether/how a quantitative approach (i.e., a deep learning-based approach) may be used to synthesize a model apt to perform specific qualitative analyses (e.g., to determine socio-historical information of a vernacular photo). Again, starting from the same point, (b) through the implementation of cross-dataset experiments, we could observe temporal shifts which may be due to intercultural influence.

Starting from point (a), the relations between quantitative methods and qualitative analyses, their potentials, and limits, represent open questions within different research communities [205, 206, 207, 208]. Due to the growth of digital and digitized data, qualitative analyses are becoming more and more expensive and difficult to apply to massive datasets, and quantitative methods seem to be the only way to deal with them [208]. Recent research in computer science has been focusing on how quantitative methods may be able to support quali-

tative analyses [208, 209, 210, 211]. However, some criticisms emerge also for quantitative methods, which may improperly apply the definition of measurement, simply matching tasks, objects, and events to numbers, according to specific rules [206, 212]. In addition, they may be misleading due to insufficient care in data collection, feature definition and processing, and adherence to the domain of interest [205]. Despite such critics, recent research in computer science has been focusing on how quantitative methods may be able to support qualitative analyses [208, 209, 210, 211]. In this case, our aim is to contribute to such debate, showing how quantitative and qualitative methods can coexist to carry out integrated analyses to evaluate more data than those usually examined in a qualitative process while adopting a well-defined theoretical foundation. In studying this kind of phenomenon, scholars usually resort to a small corpus of photos often gathered and verified adopting custom protocols [203, 204] and draw their conclusions embracing qualitative analyses approaches [213]. The adoption of qualitative methods has been so far justified by the small number of items socio-historians have at their disposal and by a general scepticism around the adoption of quantitative methods. We show how mixing qualitative and quantitative methods may overcome such difficulties. Therefore, we verify whether quantitative techniques, built resorting to results obtained from qualitative processes, may be employed to perform specific qualitative analyses, yielding a mixed qualitative-quantitative one. More in detail, we resorted to our socio-historical classification toolchain [7], focusing on the estimation of the socio-historical content [214]. In this way, we focus on the relationship between quantitative and qualitative methods considering the specific case of socio-historical analyses [8].

Moving to point (b), scholars from different fields agree in identifying family album photo collections as capable of capturing salient features regarding the evolution of local communities in space and time, representing an unparalleled source of information regarding the recent past [201, 202]. The different clothes that people wear, their haircut styles, the tools and machinery, the natural landscape, the overall environment, etc., may exhibit the culture of a given time and place. All of these visual features not only may amount to important cues to estimate the

shooting year [215] but could be the key point for observing and verifying possible intercultural influence (i.e., the adoption of different customs and habits in different epochs and countries). To this aim, we always resorted our socio-historical classification toolchain [7], this time focusing on the dating task (i.e., to estimate the date a photo was taken) and performing a more thorough analysis with respect to this task. In fact, we resorted also other datasets and models from [216, 217], previously presented in the literature, in order to carry out cross-dataset experiments, in addition to the IMAGO dataset and our classifiers. In this way, deep learning models revealed their potential not only in terms of their performance but also in terms of their possible applications to intercultural research.

Exceeding our initial expectations, such an approach has revealed its merit in terms of performance, but also in terms of the foreseeable implications for the benefit of different socio-historical researches. For these reasons, this contribution can be set at the intersection of socio-historical studies, multimedia computing, and artificial intelligence. Finally, with this work and all the related publications [7, 8, 9], we have laid the groundwork for further work, both in the field of user interface experience [10] and in the field of digital twins [12], respectively. Nevertheless, details of these works will be given in the next (Chapter 5).

The remainder of this Chapter is organized as follows. In Section 4.2 we review the state-of-the-art that falls closest to our contributions analyzing, in particular, those regarding the vernacular photographs, datasets and tasks (Section 4.2.1), and those regarding quantitative methods for qualitative analysis, potentials and limits (Section 4.2.2). In Section 4.3 the main contributions of this work are collected. In Section 4.4 we provide a description and the main characteristics of the dataset introduced and adopted in this work dwelling more on the annotation process (Section 4.4.1), the socio-historical context (Section 4.4.2), and the data class distribution (Section 4.4.3). In Section 4.5 we sketch the necessary socio-historical background in order to carry out the socio-historical analysis taken into account. Then, we present in Section 4.6 the multimedia tool designed to assist socio-historians and proceed in Section 4.7 with the method description dividing data preprocessing (Section 4.7.1), data partitioning (Section 4.7.2), model ar-

chitecture (Section 4.7.3), and training setting (Section 4.7.4). We validate the models trained on the proposed dataset to define an evaluation baseline for both the socio-historical tasks, i.e., socio-historical context classification and dating, and the different deep learning models considered, respectively in Section 4.8 (divided in Sections 4.8.1 and 4.8.2), and Section 4.9 (divided in Sections 4.9.1 and 4.9.2). In Section 4.10 we compare the classification performance of our application with the results obtained by a socio-historical scholar. Follow Section 4.11 that, considering a different perspective, explores if/how quantitative method could support qualitative analysis and Section 4.12 that reports and discusses cross-dataset experiments from an intercultural influence perspective. Finally, in Section 4.13 conclusions are drawn and possible future work directions are provided.

4.2 Related works

In this Section, we report the works, present in literature, that fall closest to our project in terms of datasets and tasks (Section 4.2.1), and terms of discussion about possible relations between quantitative methods and qualitative analyses, their potentials, and limits (Section 4.2.2).

4.2.1 Vernacular photograph, datasets and tasks

Considering the state-of-the-art in the meantime this work was carried out, and the works that fall closest in terms of datasets and tasks, only a few have so far analyzed analog collections of vernacular photographs [216, 217, 218]. For example, Ginosar et al. [216] employed a deep learning approach to analyze and date $37,921$ historical frontal-facing American high school yearbook photos taken from 1928 to 2010 [216]. Here, a convolutional neural network architecture was trained to analyze people's faces and predict the year in which a photo was taken. In addition, the authors observed gender dependency in the performance of dating models. Along the same line, Salem et al. [217] presented a dataset containing images of students taken from high school yearbooks, covering the 1950 to 2014 time span, considering $1,400$ photos per year. They resorted to convolutional neural networks

Original dataset	Type(s) of photography	Type(s) of camera	Theme	Cardinality	Period
[216, S. Ginosar et al.]	Portrait	Digital and analog	Frontal face from High school yearbook	168,055	1905 - 2013
[217, T. Salem et al.]	Portrait	Digital and analog	High school yearbook	ca 600,000	1912 - 2014
[218, E. Müller et al.]	Vernacular and landscape	Digital and analog	No specific theme	1,029,710	1930 - 1999
IMAGO collection	**Vernacular**	**Analog**	**Family albums**	**ca 80,000**	**1845 - 2009**

Table 4.1: Characteristics of existing datasets and IMAGO.

to estimate the shooting year of each image. To assess the characteristics that allow correctly classifying a picture, they considered both color and grey-scaled images containing faces, torsos (i.e., upper bodies including people's faces), and random regions from the images. The best performance was obtained considering color images portraying the torso of people. In addition, their results provide cues that human appearance is related to time. Instead, Muller et al. [218] analyzed the dating through the lenses of vernacular and landscape photos belonging to years 1930 to 1999, including at most $25,000$ pictures per year. The authors proposed different baselines relying on deep convolutional neural networks, considering the dating as both a regression and a classification task. Differently, Molina et al. [215] formulated the date estimation task as an image-retrieval one where, given a query, the retrieved images are ranked in terms of date similarity. For their study, they analyzed the same public dataset employed in [218]. In Table 4.1 the characteristics (e.g., image content, number of images, and covered time span) of the archives employed in the works described so far are reported. In most cases, only specific subsets of such archives have been analyzed using computer vision techniques. To provide a comfortable comparison, the same information regarding the IMAGO collection (i.e., the collection originating the dataset analyzed in this work) is provided in the last row of the same Table.

Other works have already investigated the digital cataloging of historical pho-

tos [219, 220, 221]. Corrin et al. in [222], for example, developed a prototype to find duplicates and tag photos depicting similar scenes in the Carnegie Mellon University Archives' General Photograph Collection. Arnold et al. in [223], instead, draw on scholarship from semiotics and visual cultural studies to develop a framework called *distant viewing*, to individuate larger patterns within a corpus that may be difficult to discern by closely studying only a small set of objects (e.g., narrative arcs in American sitcoms). Wevers et al. in [224], again, published one of the works that fall closest in scope, where the CHRONIC and the SIAMESET datasets were introduced to study the transition from illustrations to photographs in the history of Dutch newspapers.

Concluding, none of the works cited in this Section, (a) utilized pre-defined socio-historical categories as means of analysis, (b) considered a cross-dataset and intercultural perspective when approaching the dating task, and (c) the family album theme. Then, to the best of our knowledge, the present work amounts to the first contribution to investigate family album photographs classification according to the socio-historical context definitions and background (Sections 4.4.1, 4.4.2, and 4.5), and the possibility to verify intercultural influence through the dating.

4.2.2 Quantitative and qualitative, potentials and limits

How quantitative methods could support qualitative analyses, their potentials, and limits, represent open questions within different research communities [205, 206, 207, 208]. On one side, the results obtained by adopting quantitative methods could converge to those returned by qualitative ones. For example, Baumer et al. [208] compared a qualitative approach, from interpretive social science, and a quantitative one, from natural language processing, on textual data showing that these methods produced similar results. On the other side, some criticism emerges also for quantitative methods, which may improperly apply the definition of measurement [206] or they may be misleading due to insufficient care in handling data (e.g., feature selection, feature extraction) [205]. For example, Choy [212] pointed out that in social studies, throughout quantitative methodologies, different people and communities characteristics (e.g., identities, perceptions,

and beliefs) cannot be meaningfully converted to numbers or adequately explained without references to the proper context in which people live. Nevertheless, despite such critics, recent research in computer science has been focusing on how quantitative methods may be able to support qualitative analyses [208, 209, 210, 211]. For example, Radford et al. [209] highlighted that, although a variety of issues have emerged with the use of machine learning models in social data analyses, their intersection has provided critical new insights into social behavior. They stated that *"machine learning can and should become a critical piece of social science"* and *"similarly, social science should become an increasingly important part of machine learning"*. Geiger et al. [210] provided a critical analysis of a corpus of research that resorted to human annotation to produce datasets for supervised learning, considering data from Twitter. They observed the similarities between creating human-labeled datasets and content analysis, underlining the importance of utilizing high-quality training data to produce high-quality classifiers. Scheuerman et al. [211] highlighted the importance of data quality comes from, which criticizes how computer vision datasets are built, emphasizing the importance of categories structuring methods.

Concluding, starting from this debate and focusing on the concrete problem of implementing a socio-historical classification toolchain for a collection of vernacular photos, we here show how (i) adopting a well-defined theoretical foundation, (ii) through the implementation of quantitative methods, (iii) can be possible to evaluate more data than those usually examined in a qualitative analysis.

4.3 Contributions

From a family album photo collection, through socio-historical knowledge, and the application of deep learning algorithms, the principal contributions of this research (until now) amount to:

(a) The introduction of a family album photo collection [7], IMAGO, comprising over $80,000$ analog photos taken between 1845 and 2009 and belonging to ca $1,500$ families, primarily from the Emilia-Romagna and immediately

neighboring regions in Italy;

(b) A deep learning-based multimedia application to assist scholars in Social History in their cataloging work [7]. In particular, this process consists in identifying the socio-historical information of an image, i.e., its socio-historical context and shooting year, according to the definitions provided in [204]. Importantly, while the dating has been so far considered in literature [216, 217, 218], the estimation of the socio-historical context has not been yet investigated;

(b1) A thorough evaluation of the performance obtained by Convolutional Neural Network (CNN) models [225, 226, 227] trained on the IMAGO dataset for both the considered tasks, i.e., estimation of the socio-historical context and shooting year;

(b2) A comparison between the performance of the adopted CNN-based approaches and a Transformer-based ones [228, 229];

(c) The discussion of how a family album photo collection obtained through qualitative approaches has been exploited to perform a quantitative analysis [8] that mimics the qualitative one performed by socio-historical scholars;

(d) The attempt to verify possible intercultural influences by analyzing the differences in dating [9], i.e., the adoption of different customs and habits in different epochs and countries, resulting from cross-dataset experiments, in which we employ also the datasets and models from [216, 217].

(e) A comparison of the performance of the implemented deep learning models, CNN and Transformer-based, with the performance obtained from expert analysis of a socio-historian [7, 8], with the additional aim of finally assessing the validity of the proposed framework.

4.4 Dataset

The IMAGO collection, that we introduced in [7], is a project started in 2004 by socio-historical scholars to study the evolution of Social History through the lenses of family album photographs. This project produced (and continue to produce, as new photos are acquired from new incoming Bachelor students in the "Fashion Cultures and Practices" course every year) a digital collection of analog family album photos gathered and maintained by the Department of the Arts of the University of Bologna (a presentation of the project may be found at `imago.unibo.it`). Now, this collection includes more than $80,000$ photos shot between 1845 and 2009, belonging approximately to $1,500$ family albums, offering the opportunity of studying the evolution of Italian society during the twentieth century. Among these images, a total of $16,642$ received labels from their owners (bachelor students in the "Fashion Cultures and Practices" course), under the supervision of a socio-historical faculty, to focus on different socio-historical aspects. In particular, among these labels, the year in which the photo was taken and the socio-historical category the photo belongs to [204]. These $16,642$ images, from now on, will be referred to as the IMAGO dataset[1], the dataset analyzed in this work. In Fig. 4.1 are shown four exemplar images from the IMAGO dataset, which belong to different decades and represent different socio-historical contexts. These images are representative of the different characteristics that may be found in each photo (e.g., number of people, clothing, colors, and location), highlighting one of the main ones, i.e., each photograph portrays at least one person.

4.4.1 Annotation process

The annotation process represents a key point for different aspects of this work. In particular, the annotation process of the photos belonging to the IMAGO dataset followed a well-defined protocol. Firstly, with a lecture, the socio-historical background, the dataset construction goals, and the different classification categories were presented and explained to the owners of the photos. For what concern

[1]The IMAGO dataset is available upon request.

Figure 4.1: Sample images from IMAGO dataset.

the socio-historical context, it is fundamental that the different categories were produced by a qualitative coding process [230] based on sociological and historical criteria derived from different researches. Secondly, a second lecture covered the annotation problem in detail, focusing on the reliability and authenticity of sources of socio-historical materials, including the shooting year. Then, this annotation process generated two socio-historical metadata per each photo: (a) the socio-historical context and (b) the shooting year [204].

Note that this entire process highlighted the importance of interviewing the original owner of the photo. In fact, in case such person(s) were not available (e.g., old photo): (a) one could find a second-hand informed party (e.g., anyone informed of the context), alternatively (b) an attempt to infer the socio-historical information could be made by analyzing any written annotations behind the photo, and (c) whenever none of these paths led to a solution no annotation would be added and the photo would be discarded. This is because, from a socio-historical perspective, the information provided by the owner of a photograph amounts to the ground truth. It is the owner that injects into the dataset the social component along with the historical one. Therefore, the owner or a directly connected party (e.g., relatives and friends) holds the ground truth. For this reason, it is not possible to resort to just any automatic labeling services (e.g., Amazon SageMaker Ground Truth or the Google AI Platform Data Labeling Service [231, 232]) to obtain a high-quality annotation of given datasets. These elements emphasize the

quality and uniqueness of such datasets, including IMAGO. Nevertheless, such an approach is not new to the computer vision community. Other works in literature have also considered as image metadata the information provided by their owners [233, 234].

4.4.2 Socio-historical context

In the following, we provide the socio-historical context categories individuated in the IMAGO dataset [235], along with a brief explanation:

- *Work*, photos belonging to this class are mostly characterized by people sitting and/or standing in workplaces and wearing work clothes and/or gear;
- *Free-time*, includes scenes of leisure time, reconstructing, wherever possible, generational and gender differences. It also includes images representing people who make new experiences, visit far-off landmarks, expand social relationships, and interact with nature;
- *Motorization*, although often closely related to the *Free-time* category, this class has been distinguished as it includes symbolic objects such as cars and motorcycles, which represent a social and historical landmark;
- *Music*, as for the *Motorization* one, this class may also include scenes from leisure time, characterized in this case by the appearance of musical instruments or events;
- *Fashion*, as clothing, represents a mirror of the articulated intertwining of socio-economic, political, and cultural phenomena. This class is characterized by the presence of symbolic objects and clothes, such as suits, trousers, skirts, and coats;
- *Affectivity*, characterized by the presence of people (e.g., couples, friends, families, or colleagues) bound by inter-personal relationships;
- *Rites*, portraits of sacred and/or celebratory events from family lives;

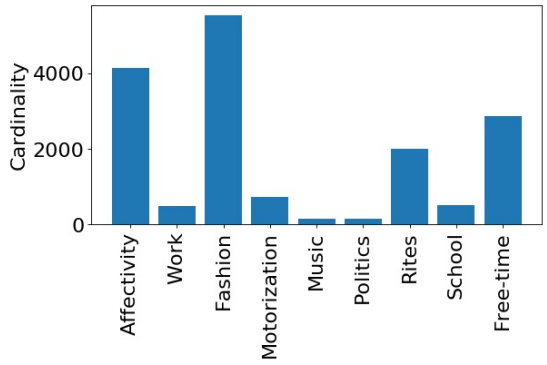
- *School*, this class includes all the photos which represent schools, often characterized by symbolic objects, such as desks and blackboards, or groups of students;

- *Politics*, this class contains photos related to political gatherings, demonstrations, and events.

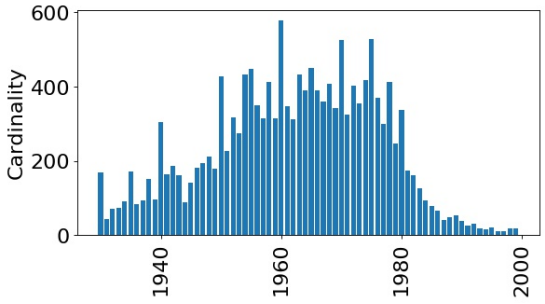
The socio-historical context categories just described employed to analyze the IMAGO dataset have been defined from a socio-historical point of view. To explain how, we here report on the rationale behind the use of two exemplar ones, *Motorization* and *Affectivity*, while a more in-depth analysis of all classes may be found in [236, 235, 214, 237]. The *Motorization* category is meant to mark an important change in people's lifestyles. We can take as an example the boom of sales for motorcycles. Such phenomena not only changed the production trend and its related economical ecosystem but also changed the social behavior of people in the area in which such a boom took place. It affected the society idea of mobility and of how people gathered together. In these terms, the motorization aspect becomes therefore fundamental for the study of Social History. On a completely different plane, instead, the *Affectivity* category regards personal feelings. This class wants to represent the changes that occurred between the affective and family relationships. For example, in the first decades of the twentieth century, family emotional relationships were considered estrangement. This phenomenon also reflects in the photographs that depict wife and husband, parents and children, brothers and sisters. Although all members of the same family, they all posed without any affectional gestures (e.g., hugs). From the second post-war on-wards, things change starting with younger people who changed poses in terms of distances, contacts, hugs, etc.

4.4.3 Data class distributions

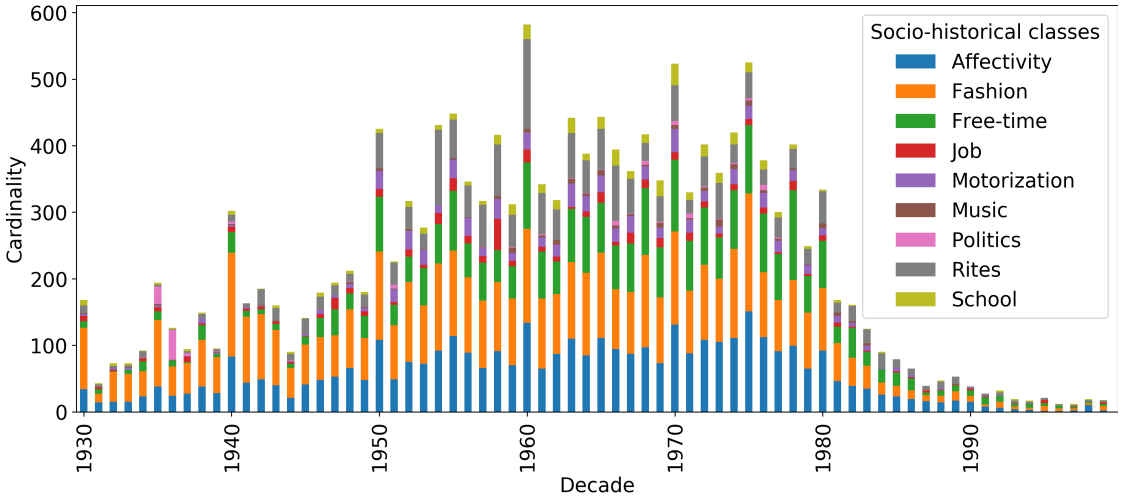
Considering the IMAGO dataset, in Fig. 4.2 we report the class distribution concerning the labels considered in this work, i.e., the socio-historical context category and the shooting year. In particular, in Fig. 4.2a is reported the distribution



(a) Socio-historical category



(b) Shooting year



(c) Socio-historical category through shooting year

Figure 4.2: IMAGO dataset class distribution.

of the socio-historical context category over the entire dataset, where it is possible to observe that the dominant classes are *Affectivity*, *Fashion*, and *Free-time*. Instead, in Fig. 4.2b is shown the number of labeled images available per shooting year in the 1930 to 1999 time frame. Among the 16, 642 labeled images, the overall available images in this interval amount to 15, 673 and, out of such time interval, the number of available images is too little to be visually represented. Here it is possible to observe that most of the considered images fall between 1950 and 1980. Then, from such plots, it is evident the unbalance that exists in terms of the num-

ber of photos both per socio-historical context category and shooting year. Finally, for the sake of completeness, Fig. 4.2c reports the image distribution considering the shooting year, but highlighting the socio-historical context category.

4.5 Socio-historical background

Now, before proceeding to explain the cataloging tool implemented in this work, it is necessary to sketch the socio-historical background [7]. In fact, no classification problem can be solved without effectively clarifying which classification categories are. This review aims at providing the basics necessary to understand how different context categories emerge in socio-historical studies. To do this, firstly (a) we delineate the main differences between traditional and social history, secondly (b) we explain how and why family photo albums fall within the areas of interest of the such field of study, and finally (c) we introduce the process that socio-historian scholars implement when cataloging a corpus of data.

Firstly, point (a), Social History amounts to an interdisciplinary field of research that combines sociological and historical methods to understand how societies have developed over time and how the past has and may influence the present [238]. In the words of Cabrera, traditional history and social history differ as follows: *Traditional History, especially classical political history, was based on the concept of the subject: the subjectivity of historical agents was rational and autonomous; the subject a preconstituted center; and, therefore, actions were caused, and fully explained, by the intentions that motivated them. Social History, on the other hand, was based on the concept of society. For social historians, subjectivity and culture are not rational creations, but representations or expressions of the social context in which the causes of actions were to be found* [239]. Such social contexts, with their own historical logic, represent the ground on which categories are constructed, to grasp the meaning and organize social reality [240]: the categories represent a complex relational network whose nature is neither subjective nor objective, but the result of a specific historical phenomenon with its own behavior. Therefore, the categories do not constitute a simple mean for transmitting

social reality, but are an active part in its definition and are called *socio-historical contexts*. The starting point of a socio-historical analysis is the space in which the interweaving between individual initiative and social coercion takes place. An attempt is usually made to explain how society works on different theoretical bases resorting to traditional oppositions: public/private, subjective/objective, ideal/ material, visible/invisible, body/conscience. Further analyses are then introduced turning to the concept of social imaginary, defined as "The way in which ordinary people imagine their social contexts which, often, does not translate into a theoretical formulation but is conveyed in images, stories and legends" [214]. In essence, any socio-historical context introduced in such analyses should describe the evolution of social history and therefore the change of sociality and of peoples' behavior in a defined space/time. To this aim, socio-historical categories are identified starting to study historical archival documents from different topics (e.g., economics, traditions, wars). Among such documents, now contemporary historians also resort to multimedia sources [241]. Out of the many multimedia sources today available, photography emerges as the one capable to cover the greatest time span so far, although, photographs have risen to the dignity of primary sources of information just in the last few decades [235].

Secondly, point (b), for the purposes of this work, socio-historical context categories have been obtained relying on the study of family album photos. This particular kind of picture originates from a well-known socio-historical abstraction that, at the same time, also represents a fundamental component of social structures: the *family* [242]. The family is, indeed, a fundamental construct in social history studies, since it embodies at once the public and the private spheres. In fact, the photos contained in the family albums can be read, on the one hand, as private visual memories of one's own history, destined to remain hidden from society and, on the other hand, as traces and signs of the collective social imaginary of a given historical period. So, family album photographs (e.g., spontaneous and/or anonymous images otherwise destined to remain hidden) depict the daily existence of their time, not considering them solely as memories, but also as a network of signs, traces, and documents that may be used to interpret the past [235].

Finally, point (c), although socio-historical contexts may emerge from the study of archival documents and family album photographs, the specific context of a specific photo may remain hard to tell. This is because, without knowing when a picture was taken and what the people there portrayed were doing, it may be impossible to associate any accurate information to the picture. In addition, accurate information in most cases may be obtained only by resorting to the knowledge of the subjects represented in the photo. For this reason, socio-historians rely on the knowledge of the main source, if available, which may be the owner of the photograph, for example. Indeed, such information could be impossible to find. In fact, when studying and cataloging a corpus of photos, no reliable source of information may be available. This problem is common for socio-historical scholars, in such a case they resort to other approaches, which may include classifying data based on a visual inspection and implementing onerous processes to reduce as much as possible the misclassifications of socio-historical features. As a relevant example consider Enns et al. [243], where the authors collected and visually analyzed and classified 355 photos related to women involved in the agriculture learning activity.

4.6 Idea of cataloging tool

Socio-historical analyses include dealing with various sources of information, systematically examining their soundness, exemplarity, meaning, and seeking for inter and intra-correlations and relationships which may help to understand what happened in the past [244]. Sources are in general not objective, but shaped by the politics, practices, and events that selectively document protest [245]. In summary, the procedure of historical inquiry implies the following steps: (i) identification and selection of sources, (ii) registration and classification for further investigation, and (iii) a critical inquiry of the collection. Starting from here, the socio-historian's work can then proceed in multiple directions. A sound socio-historical study may hence require the inspection and classification of hundreds or even thousands of documents and images [246, 247, 243]. This amounts to burdensome work which often seeks the big picture provided by large corpora of data, rather than the specific

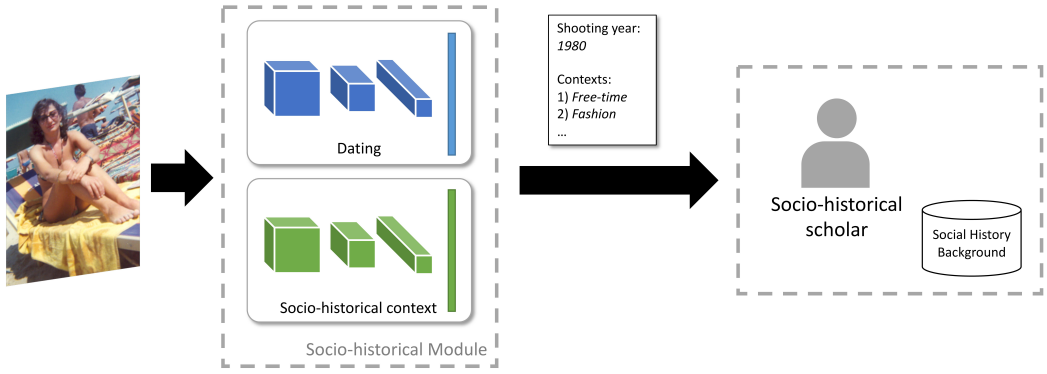


Figure 4.3: Schema of the multimedia support application for socio-historians.

information returned by a single document or image. Such type of process opens to the use of automatic tools, capable of classifying huge amounts of data in short amounts of time. This has already been discussed over two decades ago, for example in [248], where linguistic and statistical tools, that could be profitably used by historians and socio-historians in the study of events, are illustrated. Nowadays, much more can be expected thanks to the development of computing tools, capable of handling growing amounts of multimedia data originating from heterogeneous sources. This would require a holistic approach taking care of source(s): (i) digitization, (ii) accessibility through standard interfaces, and (iii) analysis with models capable of translating socio-historical tasks into computing ones.

A typical socio-historical task amounts to inferring from and subsequently applying categorical models to large corpora of data. We apply the such idea to the case of family album photos, proposing a multimedia tool capable of processing and cataloging such type of pictures. To this aim, in Fig. 4.3 the components of the proposed application are reported. The core is the Socio-Historical Module (SHM), which is composed of one or more classifiers, depending on the socio-historical tasks of interest. For the purpose of this work, such tasks have been defined on top of family album photos, originating from the IMAGO dataset (details regarding its socio-historical value are discussed in Sections 4.4 and 4.5). This dataset offered the opportunity of predicting two socio-historical information: the context and the shooting year. In brief, the SHM amounts to a tool that may

automatically label photos with the obtained predictions giving, in addition, the opportunity of confirming or correcting such estimates, when necessary, during cataloging procedures.

The classifiers that compose the SHM could be defined by exploiting different kinds of computer vision techniques. However, in the last decade, deep learning approaches have generally provided higher accuracies [249], both for the dating task [216, 217, 218] and for the analysis of historical image datasets [222, 223]. For such reasons, we also exploited such tools in the development of the SHM. In particular, inspired by the work of Salem et. al [217], we trained several classifiers considering different image regions, belonging to the same picture, selected using different criteria. We focused on the whole image and the crops enclosing the faces and the full-figures of the people there portrayed. Such patches are always present when dealing with family album photos which always include at least one person in each. To effectively estimate the value provided by such patches in terms of prediction performance, we also considered random ones. Hence, for the whole image and each of the aforementioned regions, we trained two specific single-input classifiers, one per each of the two socio-historical tasks of interest. Such classifiers are named following the analyzed patches: *full-image*, *faces*, *people*, and *random-patches*. The single-input architecture utilizes either a Convolutional Neural Network (CNN) or a Transformer as backbone and a fully connected layer for the final classification (more details in Section 4.7). Importantly, the results of such classifiers may not be comparable, as the amount of data utilized to perform a prediction varies depending on the fact that the full-image is used during testing, or parts of it (patches). This fact required establishing a different evaluation method, considering not a single face/person/random-patch, but introducing a layer that merged all of such activations into a single one per each picture. In practice, the activation vectors returned by a single-input classifier (e.g., the *face* classifier) for each region were averaged per each image in order to compute the most probable class.

Finally, we also exploited the ensemble of these models (Fig. 4.4). We resorted to such an approach as it has been successfully applied in literature [250]

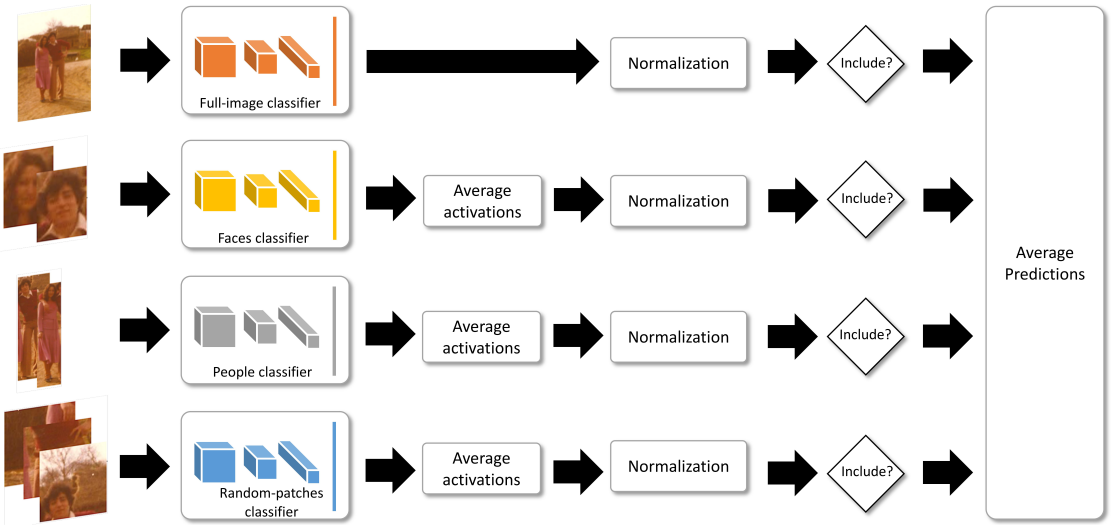


Figure 4.4: Ensemble of the different models trained on the proposed datasets; depending on the information exploited to obtain the final prediction the activations from a model may be included or not.

and did not require any additional training and tuning of hyperparameters. This kind of approach was employed, not only to exploit the averaging effect [251] but also because it helps identify which type of classifier and data provide a valid contribution at inference time. As represented in Fig. 4.4, such an approach is modular, supporting the selection of the single-input classifiers. However, since we are considering activations coming from a single image or obtained averaging across multiple regions, these may contain values at different scales. For this reason, we l2-normalized the different inputs of the ensemble to support the combination of the activation vectors coming from the *full-image*, *faces*, *people* and *random-patches* classifiers. Then, the final prediction is obtained by averaging the outputs described above and computing the most probable class.

4.7 Method

The entire IMAGO dataset, 16,642 labeled photos spanning from 1845 to 2009, was used during the analysis of the socio-historical context classification task. For

what concerns the dating task, instead, the 15,673 pictures covering the 1930 to 1999 temporal interval have been employed to avoid those years with a very limited number of samples (see Fig. 4.2b).

4.7.1 Data preprocessing

The preprocessing phase aimed principally at (i) isolating the regions of interest from each photo, and (ii) improving the quality of the images composing the dataset, resorting to different techniques.

Starting from point (i), since in this setting both faces and people could represent regions of interest to be exploited for such analysis [217, 216] (the socio-historical background is reported in Section 4.5), we created the IMAGO-FACES and the IMAGO-PEOPLE datasets, comprising over 60,000 samples each. The first is composed of individual faces, while the second is composed of single person's full-figure images. Importantly, as aforementioned in the previous Section, such patches are always present since we are dealing with photos that always include at least one person. These have been obtained by processing each image of the IMAGO dataset using the open-source implementations of YOLO-FACE and YOLO available at [252, 253], respectively. The IMAGO-FACES dataset has been constructed accounting for the number of people portrayed in a photo. In fact, by adopting a fixed-size bounding box it may be possible to lose relevant details (e.g., hairstyle) or to include pixels related to the faces of other people. To avoid such a problem, an adaptive strategy has been adopted and the size of the bounding box used to crop a face depends on the number of people portrayed in a photo: the greater the number of people, the smaller the bounding box. In this way, it was possible to extract the shoulders and the full head of a single person even when a picture portrayed tens of people. The construction of the IMAGO-PEOPLE dataset follows the same criteria employed for IMAGO-FACES, though, images can present different aspect ratios (i.e., people may be standing or sitting in photos). In Fig. 4.5 are shown some sample images taken from the IMAGO-FACES and the IMAGO-PEOPLE dataset, respectively, considering different decades and different socio-historical contexts. It is possible to appreciate

Figure 4.5: Sample of different patches: IMAGO-FACES, IMAGO-PEOPLE, and IMAGO-RANDOM.

that IMAGO-PEOPLE includes details that are not present in IMAGO-FACES (e.g., the clothing of a person). The IMAGO-FACES and IMAGO-PEOPLE were defined only to fine-tune the deep learning models for the socio-historical tasks introduced with the IMAGO dataset. So, we will not release such datasets, since their creation is technology-dependent. Indeed, in the future, algorithms or models providing more accurate bounding boxes for faces and people regions could be introduced. Finally, to study the possible usefulness of non-human features within a family album photo dataset, we created a dataset called IMAGO-RANDOM, comprising 8 randomly cropped regions, of 128×128 pixels, from each image in the IMAGO dataset (some samples are reported in Fig. 4.5). Other window sizes were also tested but returned a lower performance.

Moving to point (ii), we verified the utility of performing denoising and super resolution operations, as all the images considered in this work derive from scans of the analog prints. For denoising, we tested the neural network model from [254], and the Bilateral Filter [255]. For super resolution, we used an open-source implementation of the ESRGAN model [256], within the Image Restoration Toolbox [257]. The overall improvement obtained from adopting such strategies was revealed to be negligible, we hence opted for an analysis based on the original scans of analog photos.

4.7.2 Data partitioning

All the datasets considered, i.e., IMAGO, IMAGO-FACES, IMAGO-PEOPLE, IMAGO-RANDOM, have been partitioned as follows: 80% for training and 20% for testing. In addition, 10% of the training images is used as the validation set for hyperparameters tuning. For each image in the train set of IMAGO, the faces and the people portrayed, and the random patches are extracted and added to the corresponding dataset subset. This process is repeated also for the validation and test sets, as it guarantees that none of the training samples may end in the validation and test sets.

4.7.3 Model architecture

In this work, we started considering *CNN-based classifiers*. In particular, for the CNN single-input classifiers, we adopt a well-known architecture pre-trained on ImageNet [258]: ResNet50 [225], InceptionV3 [226] or DenseNet121 [227]. This architecture was modified replacing the top-level classifier with a new classification layer, whose structure depends on the socio-historical task (i.e., the number of output classes) and whose weights have been randomly initialized. In addition, the pre-trained convolutional layers have been specifically fine-tuned for the given input data and task. Different architectures were considered, in order to verify the independence of our dataset from the specific architecture itself. However, the results were very similar (Tables 4.2 and 4.7, Sections 4.8.1 and 4.9.1, respectively), then we decided to choose the ResNet50 as the main backbone for our analysis since it represents a good trade-off between performance and number of parameters [259]. Nevertheless, also *Transformer-based classifiers* exist. The Transformer is a deep learning architecture that relies entirely on the self-attention mechanism to draw global dependencies between input and output [260], and recent works have shown that such an approach can achieve comparable or even superior performance than CNNs [228, 229, 261]. In particular, the Vision Transformer (ViT) architecture, proposed by Dosovitskiy et al. [228], has achieved state-of-the-art performance on several computer vision benchmarks. For these reasons, we decided to exploit the

Transformer architecture and, more in detail, the ViT architecture in this work, in addition to the CNN one. For the Transformer single-input classifiers we then proceeded fine-tuning different ViT configurations (i.e., Tiny, Small, Base, and Large) varying the size of the input images (i.e., 224×224 or 384×384) and considering patches of 16×16 pixels.

4.7.4 Training setting

During the training phase of the *CNN-based classifiers*, we applied data augmentation (e.g., random crop and horizontal flip), in order to make the model less prone to overfitting. Each model has been fine-tuned using a weighted cross entropy loss to counter the unbalance in our dataset [262]. The Adam optimizer has been employed with a learning rate of $1e$-4 and a weight decay of $5e$-4. We set the batch size to 32 for the training of the *full-image* classifier and to 64 for the *faces*, *people*, and *random-patches* models. Instead, for the *Transformer-based classifiers* training, we followed the procedure reported in [228], while adopting a weighted cross-entropy loss to counter the dataset unbalance [262] and preserving the subdivision in training, validation, and test sets used in our other experiments.

These processes were adopted, respectively, for both the socio-historical context classification and dating tasks for all of the proposed datasets, i.e., IMAGO, IMAGO-FACES, IMAGO-PEOPLE, and IMAGO-RANDOM.

4.8 Socio-historical context classification results

The entire IMAGO dataset is used in our analysis for the socio-historical context classification task. The results are reported in terms of top-k accuracy, i.e., if the correct class is not the one with the highest predicted probability, but falls among the k with the highest predicted probabilities, it will be counted as correct.

Full-image classifier

	CNN-based		
Architecture	**DenseNet121**	**InceptionV3**	**ResNet-50**
#params (K)	6,963	25,130	23,526
input dim	256	299	256
Top-1	63.72	64.08	**64.35**
Top-2	83.38	83.83	**85.00**
Top-3	92.37	92.28	**92.85**
Top-4	96.54	**96.75**	96.66
Top-5	98.47	**98.53**	98.35

Table 4.2: Accuracy for the socio-historical context *full-image* classifiers considering the CNN-based models and the Top-k predicted classes (k ranging from 1 to 5).

Single-input classifiers

	ResNet50-based			
Patches	**full-image**	**faces**	**people**	**random-patches**
Top-1	**64.35**	41.30	56.54	37.35
Top-2	**85.00**	65.55	78.48	62.40
Top-3	**92.85**	82.75	89.90	80.31
Top-4	**96.66**	90.86	94.74	90.42
Top-5	**98.35**	94.98	97.42	95.35

Table 4.3: Accuracy for the socio-historical context single-input classifiers considering the ResNet50-based models and the Top-k predicted classes (k ranging from 1 to 5).

4.8.1 CNN-based classifiers

Firstly, we proceed to report on the performance obtained with the CNN single-input classifiers and with the Ensemble model. Secondly, we provide a qualitative Grad-CAM analysis of the behavior of these classifiers.

Single-input classifiers Regarding the single-input classifiers performance, in Table 4.2 are reported the results considering the CNN *full-image* classifiers, described in Section 4.7, considering the well-know architecture DenseNet121, ResNet50, and InceptionV3. It is possible to observe that the results are very similar. Then, considering the trade-off between performance and the number of parameters, we decide to choose the ResNet50 as the main backbone for subsequent analyses. In Table 4.3 are reported the results considering all the ResNet50 single-input classifiers (i.e., *full-image*, *faces*, *people*, and *random-patches*). The *full-image* classifier exhibits a higher accuracy compared to the others.

	Single-input classifiers			
Category	**full-image**	**faces**	**people**	**random-patches**
Affectivity	**64.54**	28.25	43.15	29.58
Work	**30.00**	24.00	22.00	29.00
Fashion	65.79	55.87	**67.60**	38.80
Motorization	**88.44**	17.01	51.02	29.66
Music	**40.62**	15.62	25.00	12.50
Politics	65.52	24.14	48.28	**66.67**
Rites	**71.50**	42.50	66.50	39.59
School	**60.58**	22.12	48.08	14.42
Free-time	58.09	46.09	**58.78**	51.94

Table 4.4: Single class accuracy for each socio-historical context ResNet50 single-input classifier.

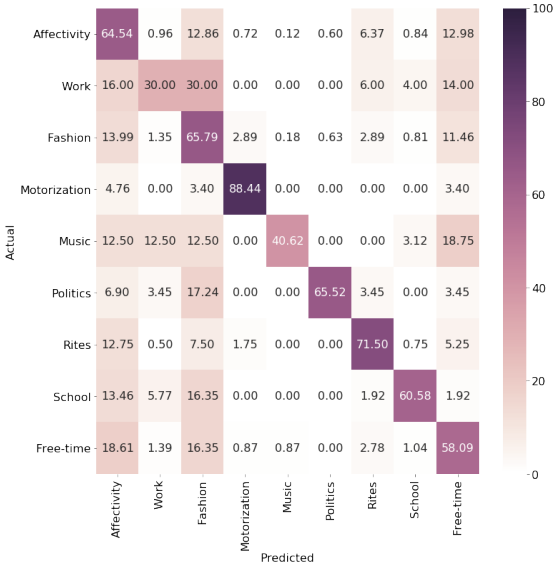


Figure 4.6: Confusion matrix for the ResNet50 *full-image* classifier.

In Table 4.4, to further investigate the reasons behind such results, we report a comparison between the accuracy of each class considering the different single-input classifiers. As it is possible to observe, the model trained on IMAGO provides the best performance for the *Motorization*, *Rites*, *Music*, *School*, *Affectivity*, and *Work* classes. This may be due to the presence of specific objects that drive the performance of the model, also considering that the model was initialized with the ImageNet pre-trained weights [258], which contains classes such as race car and car wheel. Indeed, from a socio-historical point of view, images from the classes *Rites* and *Music* could contain physical objects and/or symbols that are representative of that class (e.g., formal attires, musical instruments). Nevertheless, such objects only acquire meaning when people deal with them. However, the fact that the *full-image* classifier reached the highest accuracy for the *School*, *Affectivity* and *Work* classes means that the network has also learned to recognize the presence of groups of people (e.g., school classes, friends standing in front of a monument, mother hugging her child) and specific clothing. Despite this classifier performing

best, some peculiar results have to be discussed. For example, the *people* classifier performs slightly better for the *Fashion* and *Free-Time* categories. This is probably because the network may be focusing on people's cloth details and poses instead of exploiting specific objects and/or backgrounds that are not present in the people's crops. Exemplar areas on which the models focus, in order to classify its images, are reported in the following paragraph regarding the Grad-CAM analysis. Finally, the *Politics* class amounts to the only one for which, in terms of performance, *the random-patches* classifier is comparable to the *full-image* one. In Fig. 4.6, in addition, is shown the confusion matrix obtained with the *full-image* classifier. It is possible to observe that the classes responsible for the largest share of misclassifications are *Fashion*, *Affectivity*, and *Free-time*. This may be due to different causes. Firstly, some classes share visual elements. For example, pictures labeled with *Work* class often depict people in uniform in workplaces. These could mistakenly be classified as belonging to the *Fashion* class, as pictures in this class are characterized by people in pose wearing some particular cloth items. Another example involves the *Music* and *Free-time* classes. Indeed, the *Music* category is characterized by photos portraying people playing some instruments or taking part in some musical event. The latter, however, could be easily associated with *Free-time* photos, since they also often portray groups of people in similar environments and poses. Secondly, the IMAGO dataset is unbalanced, as reported in Fig. 4.2a, Section 4.4.3. Indeed, the most misclassified classes are also those which contain fewer samples.

Ensemble classifiers Regarding the ensemble classifiers performance, we also evaluated different ensemble classifiers obtained from the combinations of the single-input classifiers. However, such combinations did not provide any significant improvement with respect to just considering the *full-image* model. For this reason, we decide to not report here the results and, from now on, we consider the *full-image* classifier for the analysis that will follow and as the socio-historical context classifier in our application (schema in Fig. 4.3).

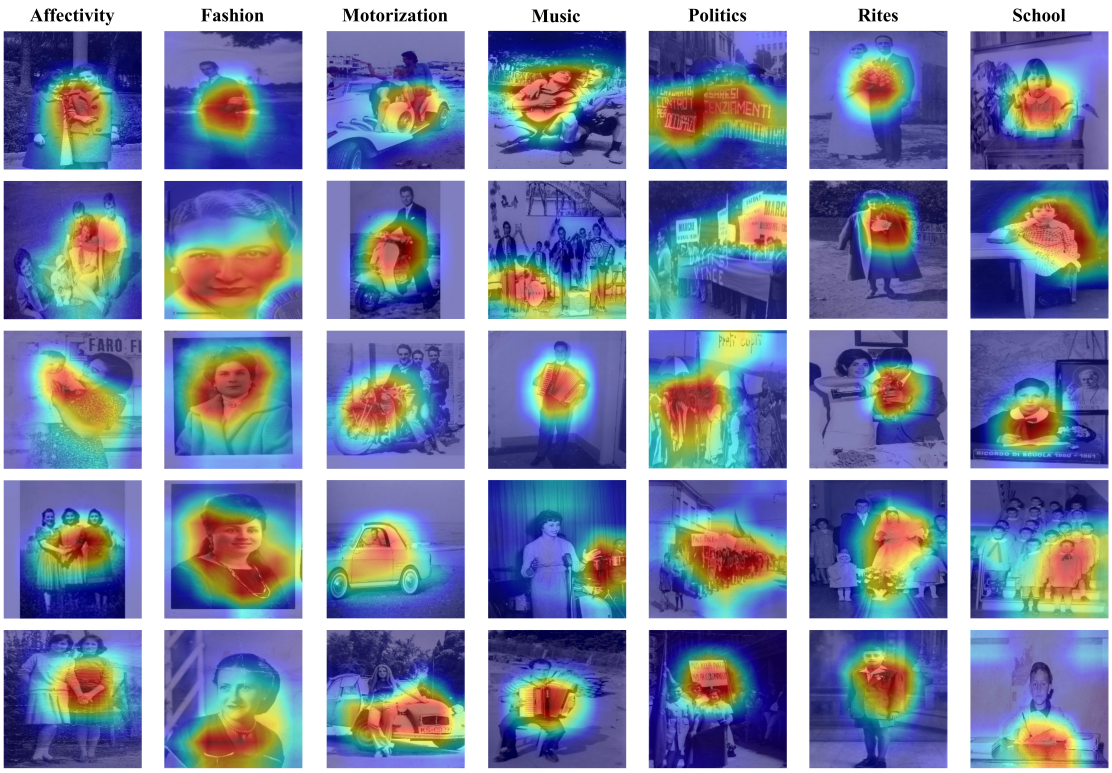


Figure 4.7: Grad-CAM analysis of socio-historical contexts of images within IMAGO, using the ResNet50 *full-image* classifier.

Grad-CAM analysis We exploited a qualitative analysis that aims at highlighting which visual cues led the classifier to associate a specific socio-historical category with a picture. To do so, we exploited the Grad-CAM algorithm [263], which delimits the areas driving the predictions performed by a deep learning model.

In Fig. 4.7 are depicted samples of correctly classified IMAGO images processed by the Grad-CAM algorithm. Each column, starting from the left, shows five exemplary images belonging to the *Affectivity*, *Fashion*, *Motorization*, *Music*, *Politics*, *Rites*, and *School* classes, respectively. Such images are representative of the regions exploited by the *full-image* classifier. More in detail, people in certain poses close to each other (e.g., hugs, holding a baby, handshakes), as shown in the first column of Fig. 4.7, are characteristic of the *Affectivity* class. Specific

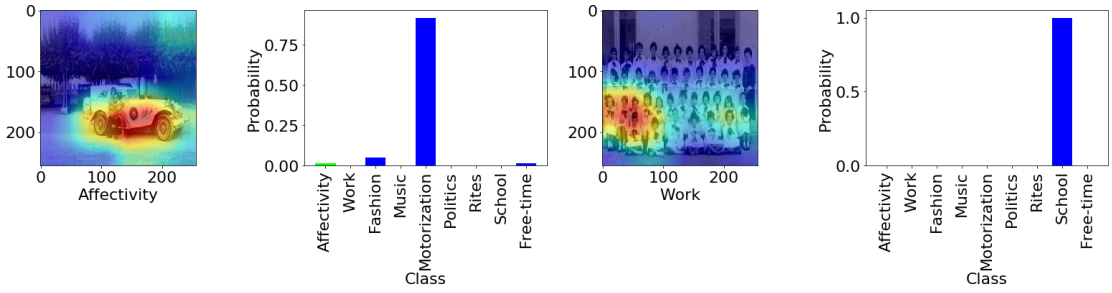


Figure 4.8: Grad-CAM examples of failure cases, considering the ResNet50 *full-image* classifier; *Affectivity* recognized as *Motorization* and *Work* recognized as *School*.

objects like earrings, necklaces, and lapels, but also particular hairstyles, are used to classify a picture as belonging to the *Fashion* class (second column of Fig. 4.7). All kinds of vehicles, as well as musical instruments, are used to recognize a given picture as a member of the *Motorization* or the *Music* classes, shown in the third and fourth columns of Fig. 4.7, respectively. The presence of a political banner is typical of pictures in the *Politics* class (fifth column of Fig. 4.7). The model also appears to individuate the objects that characterize the *Rites* class (e.g., white dress, flowers, pour a drink, cheers), as shown in the sixth column of Fig. 4.7. Finally, children wearing school uniforms, as well as school gear (e.g., books, pens, desks) are used to recognize pictures in the *School* class (last column of Fig. 4.7). As stated before, it is not surprising that the model was able to correctly classify pictures belonging to the *Motorization* and *Music* classes, as these are clearly characterized by specific objects and, more importantly, already part of the model pre-trained on ImageNet [258]. However, also for the majority of the other classes (not studied so far in the literature, to the best of our knowledge), the model seems to be able to isolate and focus on the details which distinguish them.

In Fig. 4.8 are shown, instead, some failure cases for the *full-image* classifier. From the leftmost picture and its probability histogram, it is possible to see that a photo containing a car was classified as belonging to the *Motorization* class, but the ground truth label assigned to the picture was *Affectivity* (two people standing close to each other). Instead, the rightmost picture and its corresponding

probability histogram show that a picture depicting a school class was classified as belonging to *School*, while the actual one was *Work* (a teacher is standing in the rightmost part of the picture). On one hand, such misclassifications may be traced back to the fact that the IMAGO dataset has been labeled by the owners of the pictures. The pictures thus convey such specific points of view, which may not be correctly predicted by the network. On the other hand, however, the point of view of the photo owner amounts to the ground truth, according to the methods adopted in socio-historical studies. In fact, the leftmost picture presented in Fig. 4.8 was classified as *Affectivity* since the owner of the photograph was the child of the couple there portrayed. The same phenomenon happens in the rightmost one since the one who labeled the photo was a teacher of those students. This proves the intrinsic challenge that the socio-historical classification task poses, since any classifier, including an expert socio-historian, may be subject to such kind of errors. For such reason, we further investigate this phenomenon in Section 4.10, analyzing the differences between the predictions obtained with the deep learning models and the choices made by a socio-historical scholar.

4.8.2 Transformer-based classifiers

The results obtained with the considered Transformer-based classifiers are shown and discussed following the same line adopted for the CNN-based classifier results. Firstly, we proceed to report on the performance obtained with single-input classifiers and with the Ensemble model. Secondly, we provide a qualitative Grad-CAM analysis of the behavior of these models.

Single-input and ensemble classifiers In Table 4.5 the results obtained with ViT single-input classifiers are available and contrasted with those obtained with the corresponding ResNet50 single-input classifiers, previously presented in Section 4.8.1. It is possible to observe that in most cases either ViT-Base or ViT-Large outperforms the ResNet50 while requiring a much higher number of parameters and thus increasing the complexity of the model. Instead, when a similar number of parameters is used (e.g., ViT-Small with input size 224×224), ViTs exhibit a

	Single-input classifiers								
	CNN-based	**Tranformer-based**							
Architecture	**ResNet50**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**	**ViT-Tiny**	**ViT-Small**	**ViT-Base**	**ViT-Large**
#params (K)	23,526	5,526	22,669	85,806	303,311	5,599	21,815	86,097	303,700
input dim	256	224				384			
	full-image								
Top-1	64.35	53.62	60.96	**66.24**	**67.87**	57.43	**65.13**	**68.53**	**69.19**
Top-5	98.35	96.63	97.72	**98.71**	**98.74**	97.14	97.84	**99.01**	**99.10**
	faces								
Top-1	41.30	35.58	41.23	**42.98**	**43.13**	35.61	37.21	40.64	39.43
Top-5	**94.98**	89.87	93.54	92.03	93.84	89.84	91.67	93.90	93.21
	people								
Top-1	56.54	48.42	53.23	56.08	**59.21**	46.58	51.99	**60.35**	**62.51**
Top-5	97.42	93.78	96.15	97.32	**97.45**	93.36	95.85	**98.02**	**97.69**
	random-patches								
Top-1	37.35	33.56	**40.29**	**44.09**	**43.78**	**39.06**	**38.08**	**43.72**	**44.34**
Top-5	**95.35**	86.22	93.20	93.05	95.28	92.74	91.49	92.74	93.57

Table 4.5: Comparison of single-input classifiers for socio-historical context classification, considering both ResNet50 and ViT-based models; the accuracy is reported considering the Top-1 and Top-5 predicted classes.

slightly lower performance.

In Table 4.6 we report a comparison between the accuracy of each class considering the different single-input classifiers, while in Fig. 4.9 the confusion matrix obtained with the *full-image* classifier, considering the ViT-Small architecture. Comparing these results with those reported in Table 4.4 and Fig. 4.6, it is worth noticing that ViT-Small obtains a more balanced per-class accuracy, respect to ResNet50. An additional observation regarding the application of ViT-based classifier to socio-historical tasks, considering both the socio-historical context classification and dating, is reported in the next Section 4.9.2.

As for the CNN-based models, we also considered different ensemble combinations of the ViT-based models, but no relevant improvements were detected and, for this reason, are not here reported.

	Single-input classifiers			
Category	**full-image**	**faces**	**people**	**random-patches**
Affectivity	**52.76**	30.41	32.21	11.00
Work	**55.00**	4.00	26.00	8.00
Fashion	54.24	55.32	**58.75**	51.85
Motorization	**95.24**	37.41	93.20	61.38
Music	53.12	12.50	**56.25**	9.38
Politics	**79.31**	51.72	58.62	59.26
Rites	**72.25**	44.75	62.75	40.61
School	**80.77**	49.04	63.46	58.65
Free-time	**66.09**	34.43	58.61	58.13

Table 4.6: Single class accuracy for each socio-historical context ViT-Small single-input classifier.

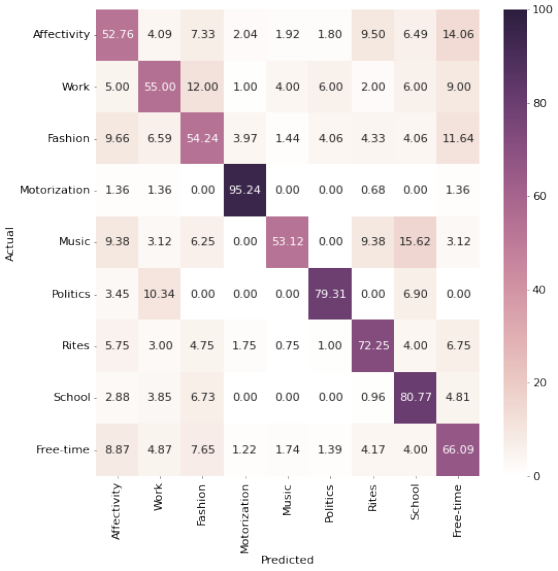


Figure 4.9: Confusion matrix for the ViT-Small *full-image* classifier.

Grad-CAM analysis In Fig. 4.10 are reported some correctly classified IMAGO pictures processed by the Grad-CAM algorithm. Per each row, five images belonging respectively to the *Affectivity*, *Motorization*, *Music*, *Politics*, and *School* classes are shown. In addition, we have images processed by the Grad-CAM algorithm applied to ViT-Small (second row) compared to the same classified with ResNet-50 (first row). Such images are representative of the characteristics that the different classifiers learned for each class. It is possible to observe that more accurate activations are obtained with ViT-Small when compared to the corresponding examples for ResNet50.

4.9 Dating results

For what concerns the dating task, as aforementioned in Section 4.7, the 15,673 images (of the 16,642 labeled pictures belong to the IMAGO dataset), covering the 1930 to 1999 temporal interval, have been employed to avoid those years with a very limited number of samples (see Fig. 4.2b). The results are expressed in

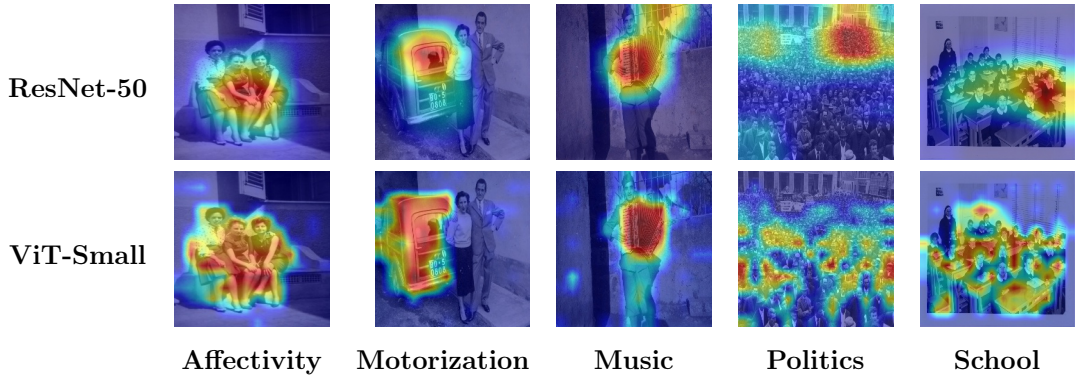


Figure 4.10: Grad-CAM analysis of socio-historical contexts of images within IMAGO, using ResNet-50 and ViT-Small.

terms of time distances, as also reported in [216, 217]. The time distance defines the tolerance accepted in predictions concerning the actual year. For example, if a photo was labeled with the year 1942 and the model returned 1937, or even 1947, this would be considered correct if the time distance is set to be equal or greater than 5, otherwise, it represents an error. In this work, model accuracies were computed considering temporal distances of 0, 5, and 10 years and have been assessed for both single-input and ensemble classifiers.

4.9.1 CNN-based classifiers

As stated for the socio-historical context classification task, also for the dating task both single-input and ensemble models are exploited.

Single-input classifiers Regarding the performance of the single-input classifiers, in Table 4.7 are reported the results considering the CNN *full-image* classifiers, described in Section 4.7, considering the well-known architecture DenseNet121, ResNet50, and InceptionV3. As for the socio-historical context classification task, also for the dating task, it is possible to appreciate that different baseline models (i.e., ResNet-50, InceptionV3, DenseNet121) return similar accuracies. Then, considering the trade-off between performance and the number of parameter, we decide to choose the ResNet50 as the main backbone for subsequent analyses,

Full-image classifier

	CNN-based		
Architecture	**DenseNet121**	**InceptionV3**	**ResNet-50**
#params (K)	7,026	25,256	23,651
input dim	256	299	256
d = 0	10.68	10.45	**11.31**
d = 5	60.77	61.38	**62.56**
d = 10	82.47	**82.82**	82.54

Table 4.7: Accuracy for the dating *full-image* classifiers considering the CNN-based models and different time distances (d = 0, d = 5, d = 10).

Single-input classifiers

	ResNet50-based			
Patches	**full-image**	**faces**	**people**	**random-patches**
d = 0	11.31	15.01	**15.77**	11.64
d = 5	**62.56**	58.09	62.40	54.26
d = 10	**82.54**	78.39	82.47	76.12

Table 4.8: Accuracy for the dating single-input classifiers considering the ResNet50-based models and different time distances (d = 0, d = 5, d = 10).

also for this task. In addition, to effectively estimate the value, in terms of prediction performance and the comparison between the potential of human (e.g., faces and people) vs. non-human features in image dating, we also considered random-patches (details in Section 4.7.1). In Table 4.8 are reported the results considering all the ResNet50 single-input classifiers, (i.e., *full-image*, *faces*, *people*, and *random-patches*). The models fine-tuned on faces and people regions achieved a higher accuracy compared to the *full-image* classifier when considering a time distance equal to 0. This is also true for the *random-patches* classifier, which performed even worse with larger time distances. These results could be explained by model averaging obtained from the ensembling of multiple image regions, as the use of more data allows controlling the uncertainty and reducing the prediction error rate [251]. Nevertheless, this increase in performance may also be due to the *faces* and *people* classifiers learning specific visual features characteristic of people's appearance (e.g., dresses, hairstyle, earrings, trousers) of given time slices. To verify whether such improvement was due to the averaging effect, we designed a specific experiment.

In particular, we considered a test subset composed of all those images containing at least $n = 8$ faces or people crops (details in Section 4.7.1). To weigh the role

	Single-input classifiers		
# of crops	**faces**	**people**	**random-patches**
1	11.70 (1.47)	11.74 (1.56)	6.35 (1.27)
2	12.88 (1.39)	14.32 (1.46)	6.97 (1.23)
3	13.46 (1.27)	15.09 (1.44)	8.01 (1.20)
4	13.87 (1.25)	15.47 (1.26)	8.15 (1.14)
5	14.19 (1.19)	15.71 (1.14)	8.16 (1.07)
6	14.40 (1.10)	15.89 (1.06)	8.42 (0.95)
7	14.58 (1.06)	16.07 (1.04)	8.47 (0.86)
8	14.82 (0.95)	15.93 (0.91)	9.00 (0.00)

Table 4.9: ResNet50 single-input classifiers averaging accuracies, along with their standard deviation, considering an increasing number of patches (faces, people, random-patches) and a time distance d = 0.

of the number of faces/people, the accuracy values were computed considering k faces/people, with k growing from 1 to n. To ensure the completeness and fairness of this experiment, $1,000$ random trials per each k faces/people/random-patches were considered. In Table 4.9 the results of such experiment have been grouped by k. It is possible to observe that, in general, averaging across multiple inputs results in a higher performance, which increases when considering the faces and people regions.

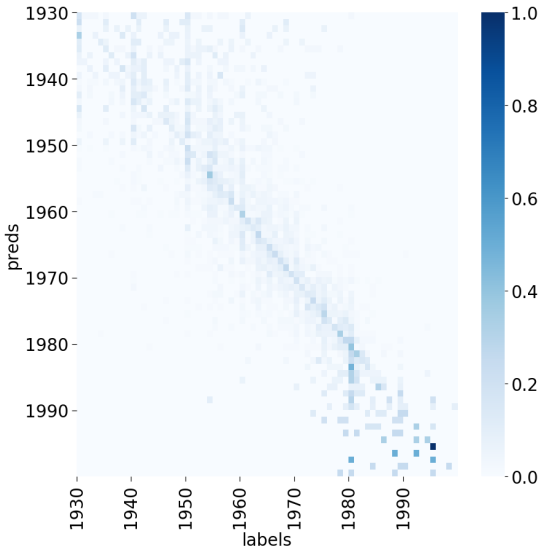
Ensemble classifiers Regarding the ensemble classifiers performance, differently from the socio-historical context classification task, an ensemble of different classifiers provides positive results for the dating task. Following the flow described in Fig. 4.4, we proceeded to evaluate different ensemble combinations, exploiting the *full-image*, *faces*, *people* and *random-patches* classifiers. Since no significative improvements were observed employing the *random-patches* classifier, for the sake of clarity, Table 4.10 only includes the results which involve the *full-image* (T), *faces* (F), and *people* (P) classifiers. It is possible to observe that the best overall

	Ensemble classifiers			
	ResNet50-based			
Patches	**T + F**	**T + P**	**F + P**	**T + F + P**
d = 0	17.14	16.79	17.91	**18.51**
d = 5	66.51	66.44	64.02	**67.53**
d = 10	85.66	84.80	83.75	**86.17**

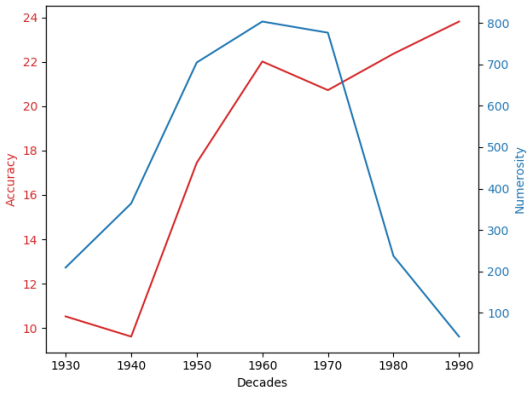
Table 4.10: Ensemble model considering different combinations of ResNet50 *full-image* (T), *faces* (F) and *people* (P) classifiers. The accuracy is reported for different time distances (d = 0, d = 5, d = 10).

performance is obtained with the ensemble combination of all these three classifiers. This shows that the model may benefit from averaging across different classifiers, as well as across multiple regions [251]. This model performs better than any single-input classifier. From now on, we consider, for all the following experiments, the model that reached the best performance which is the ensemble of the *full-image*, *faces*, and *people* classifiers. Moreover, this ensemble has been adopted in our application to estimate the shooting year of a picture.

In Fig. 4.11a, in addition, is shown the confusion matrix considering a time distance equal to 0. The diagonal structure demonstrates that the confusion mostly occurs between neighboring years, except for the initial and the final decades (this has been observed also in other works, as in [216]). The confusion created within the first 20 years may be caused by the low quality of the images and the limited number of samples representing those years. The confusion created within the last 20 years, instead, may be related to the fact that the number of images for these years is very limited (Fig. 4.2b, Section 4.4.3). Nevertheless, it is interesting to observe the information provided in Fig. 4.11b, where the model accuracy and the number of samples per decade are reported. This Figure confirms the finding exhibited by the confusion matrix, the model accuracy improves after the 50's. Fig. 4.11b also shows that, despite a reduction in terms of available samples per decade after the 80's, the performance of the model does not decrease.



(a) Confusion matrix for the dating task considering a time distance d = 0



(b) Model accuracy (red line) and number of samples (blue line) by decade for a time distance d = 0

Figure 4.11: Dating task measures for the ensemble model.

The accuracy generally improves after the 50's (also when the number of samples drops), and again this could be related to the fact that the images are of better quality than the previous decades.

Merged classifier For the dating task, regarding the multi-input classifiers, we stated an additional experiment. This decision was due to the fact that an ensemble of different single-input classifiers (i.e., *full-image*, *faces*, and *people*) provides positive results for this task. In particular, we defined the Merged classifier: a model which combines the single-input classifiers introduced before, with the aim not only to exploit different sources of information but also to learn how. Hence, a new training session was carried out as the newly introduced network was asked to learn how to perform such a combination. In particular, the pre-trained single-input classifiers were employed, but the classification layer was removed, preserving the CNN backbone as feature extractors. Adopting such architecture, the cardinality of the different extracted feature vectors depends on the number of faces/people

portrayed in an image, and the average of such feature vectors was computed to combine them with the vector obtained from the full-image. As a picture could contain more than one person, multiple IMAGO-FACES and IMAGO-PEOPLE images could stem from a single one in IMAGO (i.e., full-images). The three resulting feature vectors were linearly combined employing a weighted sum, whose weights were a set of three real scalars learned during the training phase. The final vector, resulting from the linear combination, is fed to a fully connected layer with a softmax activation function, yielding the final probability vector used for the classification. Regarding the training set, we proceeded in the same way reported in Section 4.7.4. Considering the ResNet50 architecture for the CNN-based Merged classifier, accuracy of **18.71**, **67.59**, and **86.17** are obtained for different time distance, $\mathbf{d = 0}$, $\mathbf{d = 5}$, and $\mathbf{d = 10}$, respectively. These results appear in line with the ones obtained with the ResNet50 Ensemble classifier, considering the same set of patches (i.e., considering the *full-images*, *faces*, and *people* classifiers), reported in Table 4.10.

When comparing the results of the different approaches, the Merged model improves compared to the single-input classifiers, as happens for the ensemble ones. In this case, this improvement can be explained by both the ensembling of multiple image regions and the fact that the Merged model has learned to fuse the features from different classifiers. Nevertheless, since the improvement with the Merged model was not significant with respect to the Ensemble one (considering the *full-images*, *faces*, and *people* classifiers), we decide to continue with the latter for further experiments.

Grad-CAM analysis Considering the dating task, we investigated which cues led the trained models to determine the specific year of a picture. Differently from the socio-historical task, such type of qualitative analysis may already be found in literature [216, 217]. Nevertheless, as for the socio-historical context classification task, we applied the Grad-CAM algorithm [263] to delimit the areas exploited by the deep learning models to perform the classification. In Fig. 4.12 are reported the results for some correctly classified images. In particular, each row

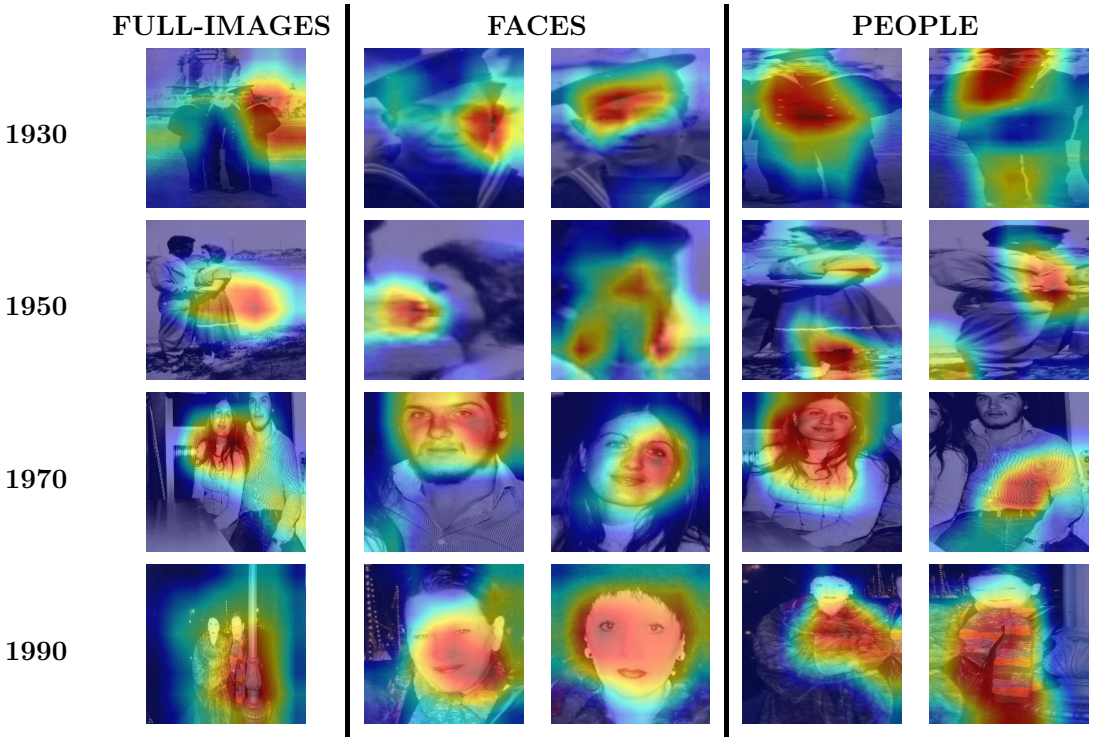


Figure 4.12: Grad-CAM analysis of estimating the shooting year of different full-images within IMAGO, and their respective IMAGO-FACES and IMAGO-PEOPLE images; samples spread over different decades.

corresponds to a specific decade and includes the Grad-CAM of an IMAGO full-image, and the two corresponding IMAGO-FACES and IMAGO-PEOPLE images, respectively. It is possible to see that the single-input classifiers focused on different regions. This may support the increased accuracy obtained in the multi-input model: different single-input classifier exploits different features. From a socio-historical perspective, these visual results may be exploited to verify whether the highlighted cues correspond to visual factors which are recognized as representative of a specific period.

4.9.2 Transformer-based classifiers

The results obtained with the considered Transformer-based classifiers regarding the dating task are shown and discussed following.

Single-input classifiers									
	CNN-based	Tranformer-based							
Architecture	ResNet50	ViT-Tiny	ViT-Small	ViT-Base	ViT-Large	ViT-Tiny	ViT-Small	ViT-Base	ViT-Large
#params (K)	23,651	5,538	21,693	85,852	303,373	5,611	21,839	86,144	303,763
input dim	256	224				384			
full-image									
d = 0	**11.31**	5.16	7.27	9.47	10.26	4.62	7.11	10.17	9.97
d = 5	**62.56**	38.85	43.40	50.72	53.44	35.21	46.37	54.11	55.74
d = 10	**82.54**	58.38	62.84	71.92	73.68	54.46	66.19	74.98	75.21
faces									
d = 0	**15.01**	3.47	5.10	6.66	7.43	4.11	4.78	6.72	7.46
d = 5	**58.09**	31.39	39.42	46.46	46.59	34.58	38.34	45.73	49.24
d = 10	**78.39**	51.21	60.77	68.51	68.58	55.32	59.85	68.01	71.70
people									
d = 0	**15.77**	4.05	4.40	7.11	7.87	4.14	4.68	7.33	7.97
d = 5	**62.40**	33.65	34.51	46.88	48.69	32.22	38.91	46.18	49.17
d = 10	**82.47**	54.21	51.56	68.90	70.24	52.42	59.40	67.81	70.04
random-patches									
d = 0	**11.64**	4.08	5.00	7.08	7.43	4.21	5.00	7.20	7.49
d = 5	**54.26**	34.08	36.05	41.82	43.39	32.83	34.11	42.17	41.89
d = 10	**76.12**	56.81	57.09	64.81	66.59	51.51	54.22	64.81	64.74

Table 4.11: Comparison of single-input classifiers for the dating, considering both ResNet50 and ViT-based models; the accuracy is reported for different time distances (d = 0, d = 5, d = 10).

Single-input and ensemble classifiers Differently from the socio-historical task (Table 4.5, Section 4.8.2), the results reported in Table 4.11 show that the ResNet50 outperforms all single-input ViT configurations for the dating task. Also, different ensemble combinations are considered, but no relevant improvements were detected and then the results are not here reported. Since both for single-input and ensemble classifiers, the ViT-based ones do not significantly outperform the ResNet50-based one, no qualitative analysis (Grad-CAM analysis) is carried out in this case.

Remark Concluding, the ViT approach exhibits divergent behaviors when applied to the socio-historical context classification and dating tasks. Why this occurred may be explained by resorting to [264], where the authors highlighted how ViT (a) incorporates more global information than ResNet at lower layers, leading to different features, and (b) strongly preserves spatial information adopting class tokens. Indeed, the inclusion of more global information at lower layers and the strong preservation of spatial information could be the reason why the socio-historical context classification task obtained a better accuracy than dating. This is also qualitatively represented by a few Grad-CAM examples reported in Fig. 4.10, Section 4.8.2: more accurate activations are obtained with ViT when compared to the corresponding examples for ResNet50. On the contrary, the dating task often requires focusing on specific local visual cues rather than on global ones, as also highlighted by Ginosaur et al. in [216].

4.10 Human vs. machine assessment

To this point, we exploited the IMAGO dataset to train the models which compose the Socio-Historical Module (SHM) (details in Section 4.6), amounting to the core of the application designed to help socio-historians in cataloging family album photographs. To assess the performance the application could attain in terms of accuracy, with respect to a human expert, we designed a specific experiment where a socio-historian was asked to categorize all the pictures in the IMAGO test set (amounting to $3,327$ images), providing both the socio-historical context category and the shooting year. On one hand, the SHM models can provide a ranking for the classes predicted for each specific photo (i.e., Top-k for the socio-historical context classification and a time interval confidence for dating). On the other hand, the socio-historian deals with the corpus of images, labeling them based on past archival and cataloging work experiences. In the following, we provide the details of the comparison with respect to the two socio-historical tasks considered in this work.

		Top-k Accuracy	
Cumulative k	**Cardinality**	**Socio-historical context module**	**Socio-historian**
1	2,147	**64.88**	54.82
1-2	3,278	**72.02**	66.53
1-2-3	3,327	**72.24**	66.93

Table 4.12: Human vs. machine for the socio-historical context classification: accuracy comparison for an increasing values of k (k ranging from 1 to 3), where k indicates the number of selections made by the socio-historian and the most probable classes returned by the model.

	Accuracy	
Time distance	**Dating module**	**Socio-historian**
d = 0	**18.51**	5.93
d = 5	**67.53**	56.36
d = 10	**86.17**	82.53

Table 4.13: Human vs. machine for the dating classification: accuracy comparison for different time distances (d = 0, d = 5, d = 10).

Socio-historical context classification task For this experiment, the socio-historical scholar was given the opportunity of selecting multiple categories per each photo. As a result of this possibility, one class was chosen for $2,147$ photos, two classes for $1,131$, and three classes for 49 images. It is interesting to point out that, although free to use as many labels as desired, no more than three have been considered at once. To make a fair comparison, we considered the k most probable classes chosen by the SHM model and compared them with the k classes selected by the socio-historian. Then, we proceeded to compute the accuracy of the socio-historian and the model, as follows. For example, if the ground truth for a photo was "*Affectivity*", the predictions provided by the application and the selections made by the socio-historian would be considered positive if both contained "*Affectivity*". Since the scholar could choose the number of categories to assign, we computed such scores cumulatively. In particular, in correspondence of *Cumulative k* with $k = 1$ (Table 4.12) a prediction is counted as positive in case it matches the ground truth. It follows that, if $k > 1$, a positive match is recorded in case one of the k predictions matches the ground truth. All the results are reported in Table 4.12. It is possible to observe that the proposed application obtained accuracy levels that surpassed those obtained by the socio-historical scholar. For example, when we consider those pictures that were tagged

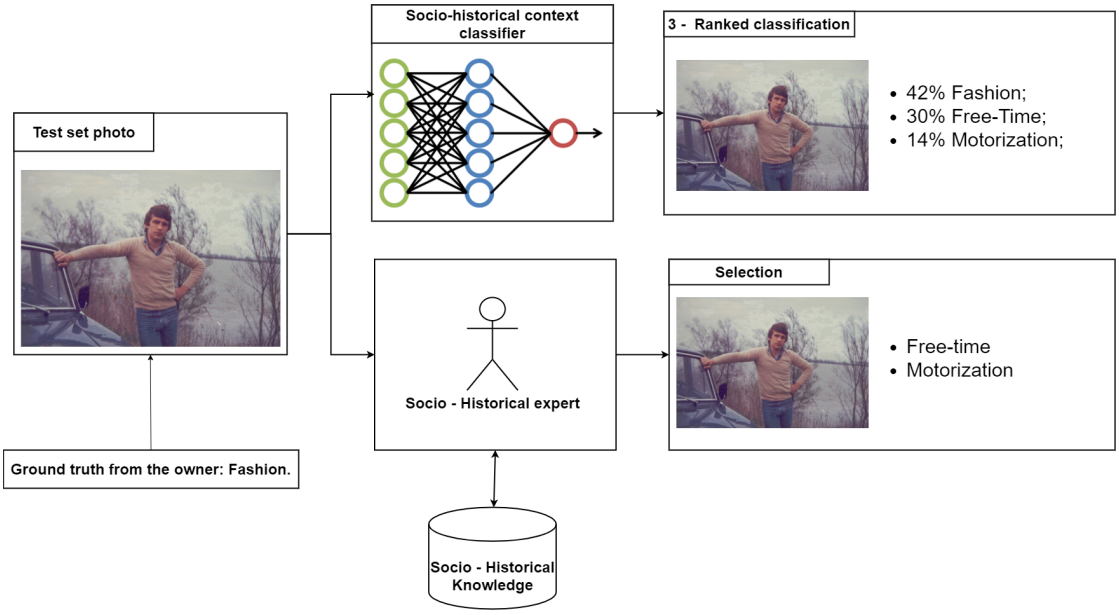


Figure 4.13: Human vs. machine experiment diagram.

with only one category by the socio-historical scholar, an accuracy of 54.82 was obtained vs. an accuracy of 64.89 for the application (+**10.07**). This occurred also when considering those pictures for which the socio-historian chose one or more classes: the application was still able to obtain higher performance. In Fig. 4.13 we show a representative example of a case where the model predicts the correct label, unlike the socio-historian. Here, the socio-historian fails at recognizing a particular detail that only the owner could have known (the subject of the photo is posing wearing a particular outfit), while on the contrary the model correctly classified this image.

Dating task Taking into account the dating task, the socio-historian labeled all the pictures belonging to the test set assigning a year in the 1930-1999 time span. The results are reported in Table 4.13. The dating module performed better than the socio-historian considering the specific picture shooting year (+**12.58** in accuracy). The difference in performance decreases when a higher time distance is considered, arriving at +**3.64** when the time distance is equal to 10.

4.11 Quantitative methods for qualitative analyses

Changing perspective (point of view), we state into the debate concerning how the integration of quantitative and qualitative methods should occur by experimenting if a quantitative approach, such as a deep learning-based classifier, may be used to synthesize a model apt to perform specific qualitative analyses regarding socio-historical aspects [7, 8].

In particular, we approached the concrete problem of implementing a socio-historical classification toolchain for a collection of vernacular photos: the IMAGO dataset. Firstly, we individuated a corpus of vernacular photographs (Section 4.4). Secondly, we involved the people included in the photos in the annotation process of a subset of the corpus of data (Section 4.4.1). In particular, for the socio-historical context classification task, we resorted to existing socio-historical context categories derived from previous qualitative studies (Section 4.4.2). Important to highlight that the well-defined annotation process amounted to the step which let us build quantitative methods to perform an analysis that typically demands qualitative ones: in brief, classify a photo according to the given socio-historical context categories and the available time span for the shooting year. Thirdly, we fine-tuned and deployed existing deep learning models to classify the entire corpus of data (Section 4.7). Finally, we compared the results obtained with our deep learning-based approach to the ones obtained by a socio-historian (Section 4.10). Hence, we focus on the relationship between quantitative and qualitative methods considering the specific case of socio-historical analyses. The results of such assessment proved that quantitative methods could not only speed up the cataloging processes but also support socio-historians in carrying out qualitative analyses of complex or large catalogs of visual information. Clearly, this is only one step in the direction of exploiting quantitative models to support qualitative analysis, which in this case may take into account all the processes involved in the complex socio-historical domain.

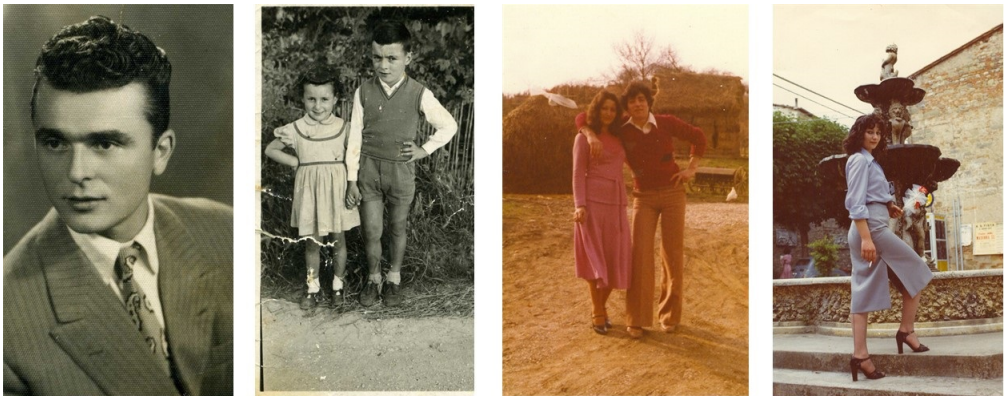
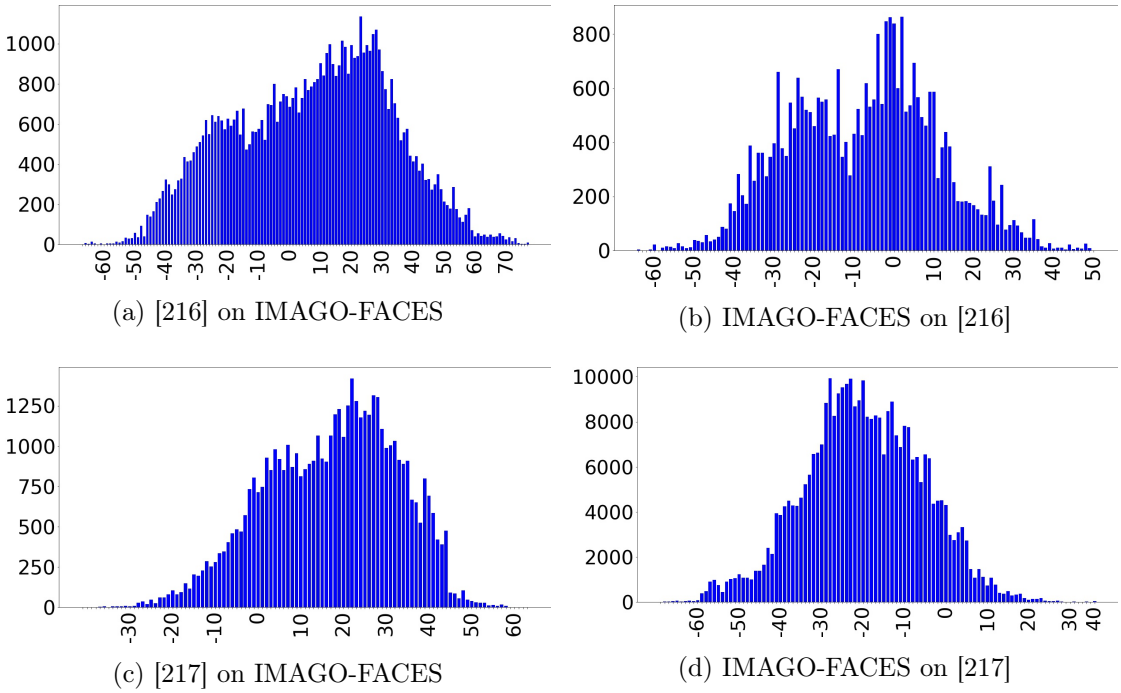


Figure 4.14: Sample images from IMAGO dataset.

4.12 Evidence of intercultural influence exploiting a cross-dataset study?

Thanks to the IMAGO dataset (more details in Section 4.4 and some exemplar images in Fig. 4.14), it was possible to apply different deep learning-based architectures to classify images belonging to family photo albums without any other sources of information, respect to their socio-historical context category and/or shooting year (results in Sections 4.8 and 4.9) [7]. Nevertheless, in this Section we want to analyze another socio-historical aspect starting from the following question: "Is it possible, exploiting the available instruments (e.g., the IMAGO dataset, implemented deep learning-based classifiers), to observe temporal shifts which may be due to known intercultural influences [9]?"

More in detail, to observe and examine the effects of possible intercultural influences (i.e., the adoption of different customs and habits in different epochs and countries) we decided to carry out a cross-dataset study, which also involves models previously presented in the literature. Then, we started to consider the datasets reported in [216, 217, 218, 215]. While [218, 215] included vernacular photos in heterogeneous settings and countries, where often no people are portrayed, [216, 217] analyzed American datasets comprising people's faces and torsos. Although such datasets do not include family album photos, they share some common traits with IMAGO: people in pictures are often in pose and dressed for



(a) [216] on IMAGO-FACES

(b) IMAGO-FACES on [216]

(c) [217] on IMAGO-FACES

(d) IMAGO-FACES on [217]

Figure 4.15: Cross-dataset experiments error distributions.

a specific occasion. In addition, it is possible to extract what characterizes all of them: people's faces and torsos. This allowed us to perform a cross-dataset comparison considering the models trained on the IMAGO-FACES and IMAGO-PEOPLE and the models trained to exploit the datasets introduced in [216, 217], switching the considered evaluating datasets. Firstly, to do this, we fine-tuned the architectures used in [216, 217] following the procedures described in their experimental sections. The dataset introduced in [216] considers people's faces, while the one introduced in [217] offers both people's faces and torsos. Secondly, we evaluated these models on the IMAGO dataset. Vice versa, the *faces* and *people* classifiers, presented in this work [7], have been evaluated on the corresponding regions in the datasets from [216, 217]. For a fair evaluation, the experiments were carried out on the 1930-1999 time span for the [216] vs. IMAGO comparison, while considering the 1950-1999 for the [217] vs. IMAGO one, respectively. Then, we collected the error between the predicted and the actual year per each picture.

The error distributions for the cross-dataset experiments involving face images are reported in Fig. 4.15. In Figs. 4.15a and 4.15c the error distributions shifted towards positive values while in Figs. 4.15b and 4.15d towards negative ones. The models built on top of American datasets [216, 217] applied to IMAGO-FACES tend to overestimate the image shooting year while the opposite phenomenon (underestimation) occurs when the model presented in this work is applied to both [216, 217]. The same phenomenon appeared considering people's torsos. This fact could be due to different reasons. The images contained within the considered datasets have been acquired from different places and locations, using different cameras and scanning devices, leading to what is defined as the problem of dataset shift. However, there is another dimension to consider, the effect of intercultural influences. Indeed, during the second half of the 1900 people's appearance from the USA and Italy were influenced by each other [265, 266]. Then, the obtained results, even if not confirmatory, provide clues about possible intercultural influences as the model trained with Italian pictures underestimates the American ones, while the model trained with Americans overestimates the Italian ones. These results are not final, but certainly motivate further investigations on this topic: deep learning models revealed their potential not only in terms of their performance but also in terms of their possible applications to intercultural influence research.

4.13 Conclusions and future works

In this work, we proposed a multimedia application to assist socio-historians in cataloging family album photos: the Socio-Historical Module (SHM). We presented the IMAGO dataset composed of photos belonging to family albums, representing a source of socio-historical knowledge. The dataset amounts to $16,642$ pictures, each of which is labeled with its socio-historical metadata: shooting year and context. We trained and tested single-input, ensemble and merged deep learning models, carrying out a comparative analysis considering Convolutional Neural Network and Transformer-based classifiers. The results showed that the Transformed-based approach could be promising also for socio-historical analysis.

This consists in identifying the sociological and historical context of a picture, according to the definitions provided by socio-historical scholars [204]. We proceeded to compare the performance of our application with the performance of a socio-historian. The results of such assessment proved that our application could speed up cataloging processes, with no loss of accuracy when compared to the performance of a human expert, thus providing important support to socio-historians. To the best of our knowledge, this is the first work addressing the socio-historical context classification. In addition, focusing on this work from another point of view, this work could be also positioned into the debate concerning how the integration of quantitative and qualitative should occur, concentrating on how quantitative methods (deep learning-based classifiers) may support qualitative analysis (socio-historical tasks). From this perspective, the results obtained in this work proved that quantitative methods could not only speed up the cataloging processes but also support socio-historians in carrying out qualitative analyses of large image collections. Last but not least, we adopted the implemented models to search for cues of intercultural influences through cross-dataset experiments. We evaluated the models trained on IMAGO-FACES and IMAGO-PEOPLE images and the classifiers trained on the datasets exposed in [216, 217], following a cross-dataset configuration. The dating error distributions exhibited an interesting symmetry that motivates further experiments.

Clearly, this only represents a step in the direction of creating a holistic approach to the socio-historic cataloging problem and exploiting quantitative models to support qualitative analysis, as many are the involved processes and sources of information. For example, in this specific case, the models were trained to utilize an unbalanced dataset and consider image regions that often included non-relevant information for classification purposes (e.g., background). In addition, when focusing on the socio-historical classification or dating, scholars perform analyses that resort at once to different sources of information (e.g., newspapers, magazines, archival documents), as well as to traces belonging to the same historical period. These represent three of the most relevant limits for this work. Hence, further investigations in this domain may consider: (a) larger amounts and more

balanced sets of data, (b) a better segmentation of the relevant areas of the images, and (c) the implementation of a multi-modal approach, capable of including also other sources of information and data formats. For what concerns the first point, the availability of larger datasets could surely improve the models discriminative power, also reducing possible unbalance problems. Regarding the second, the use of segmentation models may benefit the individuation of more relevant regions. For the third one, multi-modal learning appears as the approach that may best replicate the comprehensive approach normally adopted by socio-historians during cataloging processes. Indeed, exploiting knowledge from historical archival documents (and other sources) could improve the general cataloging and analysis effort. For example, knowing how people dressed during a specific period might improve the classification for both the socio-historical context and dating. Such a path, although complex, may not be impossible to follow. Indeed, recent natural language processing solutions can provide discriminative features that could be exploited in our models to improve the overall performance [267].

Chapter 5

Machine learning to improve interface user experience

Considering AR, and MAR interface applications in different contexts

5.1 Introduction

The advancements in terms of networking, image resolution, computer vision, and mobile cloud computing performance, together with the improvements of mobile devices (i.e., tablets and smartphones), are transforming Augmented Reality (AR) and Mobile Augmented Reality (MAR) into a technology which may be put to good use in a variety of contexts, expanding from an only academic or highly specialized technology to an everyday one [268, 221, 30]. Nevertheless, arriving to overlay rendered virtual annotations on top of the camera view of the real world (e.g., photo albums, product labels) often requires not only AR/MAR interface implementation but also intensive use of computer vision paradigms for object recognition and tracking, data availability, database structuring, and machine learning models.

In this work, we propose different AR/MAR applications and systems in different contexts, from cultural heritage through family photo albums, to the wine domain with a wine recognition process, passing from an artisan work environment

like the locksmith one. These applications/system proposals were also published and submitted in [10, 12, 13] and [14], respectively.

Collaborative Photo Environment (CPE) Inside the cultural heritage field, photographs amount to a prominent example of materials as a testament to the past. The photos, yielding information and clues about what happened in different situations, represent a unique chance to revive old memories about affections, relatives, friends, special events, etc. Moreover, at the time of social distancing, since the advent of the Covid-19 pandemic forced people to stay at home, away from places of interest, families and friends, photographs may represent a link between people and a distraction from worries and fears. Starting from this, and from a previous work [269], we propose an AR application as a digital application that may bring people together and support the exploration of the content of family photo albums with the aid of machine learning models [10].

Mobile Key Recognition (MKR) Considering an artisan's craft rooted back into the past, the key locksmithing, we want to show how MAR capabilities may simplify and ameliorate the performances of such ancient trade. Key locksmiths are professional figures existing since the 18th century, and their main tasks are designing and repairing locks and keys. Nowadays, key locksmiths already put to good use advanced tools on the job (e.g., lock by-pass), but one of the slowest parts of their job remains the key type recognition. This is a known problem in the industrial field, in fact, highly specialized hardware, namely key readers, is employed to reduce the recognition time. This project aims at introducing the requirements posed by such craft, proving that also a MAR-based application may be implemented to support and speed the execution of the key type recognition process. In particular, a synergistic approach is adopted, putting together the use of AR and machine learning paradigms. Furthermore, this could impact the everyday lives of ordinary citizens, since it poses the bases of remote locksmithing activities thinking, for example, of a mobile application in which the user exploits one of her/his cameras to identify and request a copy of a key, independently from its location [12].

Augmented Wine Recognition (AWR) Since in some domains the lack of reference images may be particularly disruptive for a recognition task, we here present a system that does not require any reference image to perform a reading and recognition system. In particular, we set in the wine domain. Considering wine bottles, in fact, labels may not be available because (a) the wineries periodically change them, and (b) specific bottles may belong to the long tail, making label retrieval difficult or even impossible. For these reasons, we decided to develop a wine recognition system based on a textual database, instead of an image one, exploiting a machine learning-based OCR and implementing a custom search algorithm. In addition, to improve the user experience and to reach better usability, we decided to support this recognition system with an AR interface [13, 14].

The rest of this Chapter is organized as follows. In Section 5.2 we review the state-of-the-art that falls closest to our work, considering their specific application domains, while in Section 5.3 the main contributions of the proposed projects are grouped. Then, respectively in Sections 5.4, 5.5, and 5.6, the CPE, MKR, and AWR applications/systems are described in detail. Finally, in Section 5.7 general conclusions are stated.

5.2 Related works

In this Section, we report the works, present in the literature, in line with our projects. For the sake of clarity, we proceed separately for the three proposed applications/systems: the Collaborative Photo Environment, Mobile Key Recognition, and Augmented Wine Recognition, respectively.

Collaborative Photo Environment Many different researchers worked on the known problem with materials belonging to the cultural heritage field, which amounts to its digitization and archiving, in order to make it available in an easy and portable way. In [221], for example, the authors posed particular attention to this topic, exploring the opportunities and the criticalities which emerge

with the use of computational systems to preserve cultural resources and local traditions (i.e., Bolognese tortellini food-making). In OmniArt [270], however, the authors digitized the dataset belonging to a museum, labeling each artwork with its author(s), period, gender, and style. Another example may be found in The Newspaper Navigator Dataset [271], where the authors described the digitization of which over 16 million pages of historic American newspapers, containing not only meta-data related to their textual contents but also the spatial regions of interest and their semantic meaning. Such kinds of datasets are not only useful from an archiving point of view but also may be exploited to increase the corpus of knowledge as a unique source to learn and produce knowledge about unknown material. In addition, it is possible to identify a clear workflow in such works: (i) digitize specific cultural heritage assets, (ii) build a dataset, and (iii) share it with the world. Nevertheless, it is also possible to find research projects that have focused only on the sharing phase. In [272], for example, the authors explored the possibility of using the camera feed to live stream artistic performances or cultural traditions and customs. Inspired by such works, in [269] the authors considered a cultural heritage dataset composed by the digitization of analog family album photos, labeled by their date and socio-historical context category (e.g., free-time, school, rites) by expert socio-historians, namely IMAGO, and the socio-historical classifiers trained on it and able to predict, respectively, the socio-historical context and the shooting year of a family album photo [7] (details in Chapter 4, where they are presented), as a starting point to an AR application. As a direct continuation, we improve this work, starting from the same workflow: a user, while browsing through a photo album, wears a head-mounted display, which captures and labels the images he/she is watching, overlapping the result in his/her view. In particular, we decided to follow this path, adding to our application the chance to share the augmented user's view with anyone, through a live-streaming system. In addition, to evaluate the effectiveness of such an AR system, we also asked a group of participants to answer some questions regarding their experience.

Mobile Key Recognition Automation tasks continuously grow in many industrial working fields, aiming to improve work efficiency and reduce the work load [273]. Indeed, as proved by the authors of [30], industrial workers, who often carry out repetitive actions (e.g., car assembly pipeline), may now benefit from the exploitation of MAR paradigms. However, it may be more difficult to realize effective MAR systems for those professionals who operate in environments characterized by a multitude of tasks. This is, for example, the case of craftsmen who often, being small business owners, are not able to benefit from standardized work protocols (e.g., locksmiths, carpenters, potters and glass-makers). Only a very limited amount of work has so far considered a such type of settings, from the point of view of digital technologies. Among these, AR systems have been envisioned as apt to encode artisans' knowledge, providing a possible response to the growing need of preserving their skills and passing them on to future generations [274, 275]. Nevertheless, the development of mobile devices, networks, and computer vision capabilities fosters the chance of providing low-cost and practical solutions for the benefit of artisans' tasks. For this reason, considering the key locksmithing, we propose a flexible MAR system capable of recognizing (and encoding) the key types to ease the artisan's everyday work.

Augmented Wine Recognition Practical product identification is often performed with barcodes or QR codes, although AR applications also rely on image recognition [276, 29, 277]. Identifying and exploiting visual cues in the photos of food products has appeared in various research contributions [278, 279, 280, 28]. Nevertheless, we will focus on the research works close to wine label recognition for AR applications, considering both industry and academic scenarios. As an online service, WineEngine recognizes wine labels [281], and using the service requires adding reference label images to the WineEngine collection. Moving to a commercial application available on mobile app stores, Living Wine Labels [282] "gives life" to the wine bottle telling stories and showing 3D content in AR, by framing the front label of one of the eleven brands it supports. Vivino, instead, is the most downloaded app with a community comprising 20 million users around the

globe [283, 284], and allows to access different wine evaluations, explore selections of wines, scan the wine bottle front labels or restaurant wine lists to find out more about the wines. Vivino does not provide an AR interface but implements an image retrieval approach resorting to the Vuforia Cloud Recognition service that compares incoming scans uploaded by the app to the stored front label images to discover the closest match [284]. Given the high number of downloads and the large community, we used Vivino to perform an experiment that further supports the rationale of our work. In particular, we tested the app on a random selection of 60 bottles found at a local supermarket: 47 were correctly recognized (78% accuracy) in 2.05 s on average (0.65 s of standard deviation). In addition, also academic contributions have followed image retrieval-based approaches. In [285] the authors implemented a front-label recognition method computing SURF key points and label descriptors and comparing such descriptors to precomputed ones in a label database to search for a match. Similar approaches may be found in [286, 287]. In [288] the authors proposed a CNN-SIFT framework for wine label retrieval, where a trained CNN model recognizes the manufacturer to narrow the search range, while a SIFT descriptor empowered with RANSAC and TF-IDF mechanisms matches the final sub-brand. In [289] the authors presented an AR system running on a Microsoft HoloLens, making use of the Vuforia SDK to recognize markers attached to the wine bottles and to display information concerning those bottles [290]. It is also possible to find other approaches in the literature, although these do not lead to a complete solution to wine recognition. In [291], for example, the authors concentrated on a preliminary step, a region of interest extraction method (GrabCut algorithm) for front labels, that may serve subsequent ones such as image analysis, recognition, and retrieval. In [292], the authors implemented an OCR-based solution to read serial numbers from wine labels to provide counterfeit prevention and brand protection. Differently, the AWR system does not rely on a label image database. The proposed approach relies, instead, on the information reported on a bottle label that uniquely characterizes each specific wine.

5.3 Contributions

Considering different AR/MAR systems, for different applications, in different contexts, the principal contributions of this work are reported in the following.

Collaborative Photo Environment A Collaborative Photo Environment (CPE) system [10] that, starting from a previous work [269], includes:

(a) The training of the YOLO architecture to recognize and crop images from family photo albums;

(b) The implementation of the AR system that pipelines the HoloLens 2 user's view capture and subsequently augments it with the information drawn by the YOLO and IMAGO deep learning models [7];

(c) The devise of a simple service to share the HoloLens 2 user's view to anyone from any kind of device;

(d) The evaluation and the validation of the proposed system through a simple assessment model, asking a group of people to provide their comments (with a survey) regarding the use of our prototype.

Mobile Key Recognition A Mobile Key Recognition (MKR) application proposal [12] that, since task automation continuously grows in many industrial working fields, it may be more difficult to realize systems for artisan professional figures. We decided to focus on how MAR technologies could support (as a low-cost and practical solution) the craftsmen professional figure of the key locksmith that, to the best of our knowledge, has not been considered so far for this research. In particular, the proposed application is a MAR application, characterized by: (a) a custom AR-guided interface for an easy key type recognition and visualization, (b) a client-server paradigm, (c) the implementation of an Optical Character Recognition (OCR) module, and (d) the use of machine learning algorithms to predict the most-probable key type based on its visual features.

Augmented Wine Recognition An Augmented Wine Recognition (AWR) system that does not resort to an image match mechanism. Such a result comes considering the text on the label that contains the relevant information characterizing a wine (usually the back label), and taking advantage of the extensive tradition existing in this domain, which translates into specific regulations on how such information is printed on such a label. In particular, a deep learning-based OCR module and a custom search algorithm were implemented. In addition, an online demonstration[1] completes the presentation of the AWR system proposed.

5.4 Revive family photo albums through an AR collaborative environment

Setting in the cultural heritage field, we here present a system to revive family photo albums through an Augmented Reality (AR) collaborative environment. In particular, the contents are organized as follows: in Section 5.4.1 the domain knowledge necessary to explain our project is reported [269, 10]. In Section 5.4.2 we proceed to introduce our system, while in Section 5.4.3 we report its details concerning the design and implementation. In Sections 5.4.4 and 5.4.5, instead, we describe the assessment model, reporting the experimental setting and evaluation, respectively. Then, in Section 5.4.6 we analyze the obtained results. In addition, in Section 5.4.7 we want to highlight how the collaborative environment presented could be nowadays strictly related to the concept of digital twins [11]. Finally, in Section 5.4.8 we conclude by providing a discussion about the overall project and possible future works in the same research direction.

5.4.1 Domain knowledge

Historical and analog photos provide an unrepeatable chance to revive old memories about social events, affections, relatives, friends, special events, etc. During the 20th century, people printed and collected such kinds of pictures in

[1] Accessible at `https://tinyurl.com/2p82vmc8`.

photo albums, namely family photo albums. Nevertheless, photos belonging to an album often lack some socio-historical information (e.g., shooting year). Follow this line, and starting from the digitization and cataloging system introduced, and explained in detail in Chapter 4, in this project we will consider the date (in particular, the shooting year) and the socio-historical context (Section 4.4.2, Chapter 4) as socio-historical information of interest. For the sake of clarity, avoiding unnecessary repetitions, we directly refer to Chapter 4, with particular attention to Section 4.5, for additional details regarding family album photos and their role in the social history.

5.4.2 Collaborative Photo Environment

With this project, we aim at giving the possibility to revive family photo albums through an AR collaborative system, the Collaborative Photo Environment (CPE). In particular, this system will be composed of different components, coming from a previous project [269], among which: the HoloLens 2 [293] as the wearable device, AR paradigms to implement our interface, and deep learning algorithms to catalog the pictures observed by the user. With respect to the previous one, the proposed system includes: (a) the chance to share with remote users the HoloLens 2 scene view, and (b) a more performing detecting and cataloging process, exploiting a well-known object detector, YOLO [294]. All these elements will be explained, in detail, in the following.

5.4.3 CPE - Design and implementation

As previously stated, this project aimed at extending the work introduced in [269], concentrating on improving the detection performance and providing an authentic experience of sharing family memories exploiting AR and deep learning techniques. To reach these goals, a custom AR system that comprehends an HoloLens 2 interface, and deep learning processing was designed. Such steps are visually represented in Figs. 5.1 and 5.2. Additional details concerning the components of the proposed system will be provided in the following.

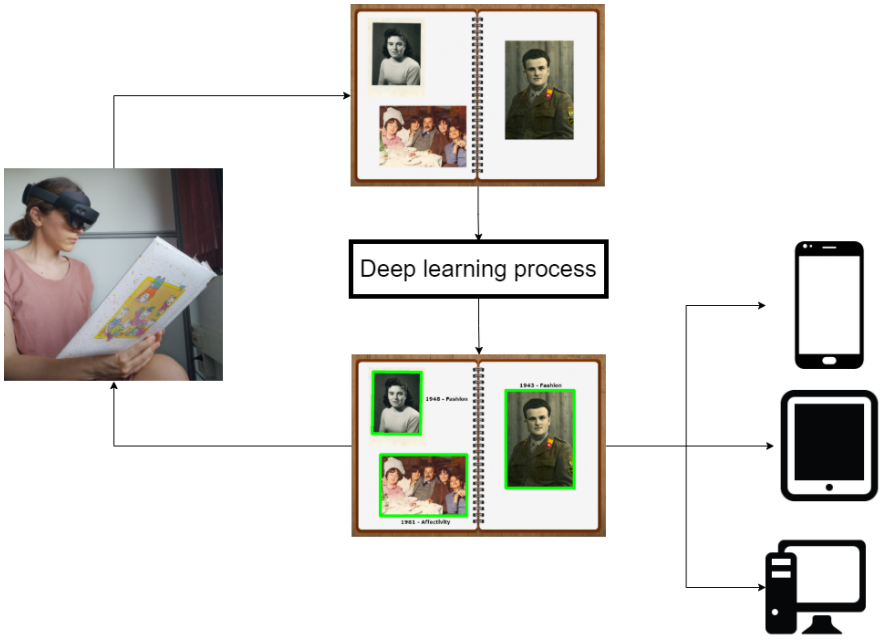


Figure 5.1: HoloLens 2 interface architecture.

AR interface and sharing As shown in Fig. 5.1, we envisioned an application for the HoloLens 2 device. In particular, the application sends all of the frames within the user's view to the deep learning models which, in case one or more pictures are detected, provides the bounding box(es) and the label(s) that can be then visualized in AR. Then, such information is utilized to augment the visualization of the family album photographs by resorting to the HoloLens 2 interface. In addition, the application supports the sharing of the augmented HoloLens 2 user's view to other devices, e.g., smartphones, tablets, and computers.

Photo detection With respect to a previous work [269], the detection module of the photos implemented in this system has been improved. The previous work, in fact, resorted to a classical computer vision pipeline to implement the task of recognizing the area in which family album photos were located. This pipeline was composed of stacking pre-processing image algorithms (i.e., bilateral filtering), edge-detection (i.e., canny edge detector), and contour-detection ones (i.e., Sobel). However, a more recent trend amounts to the exploitation of the performances

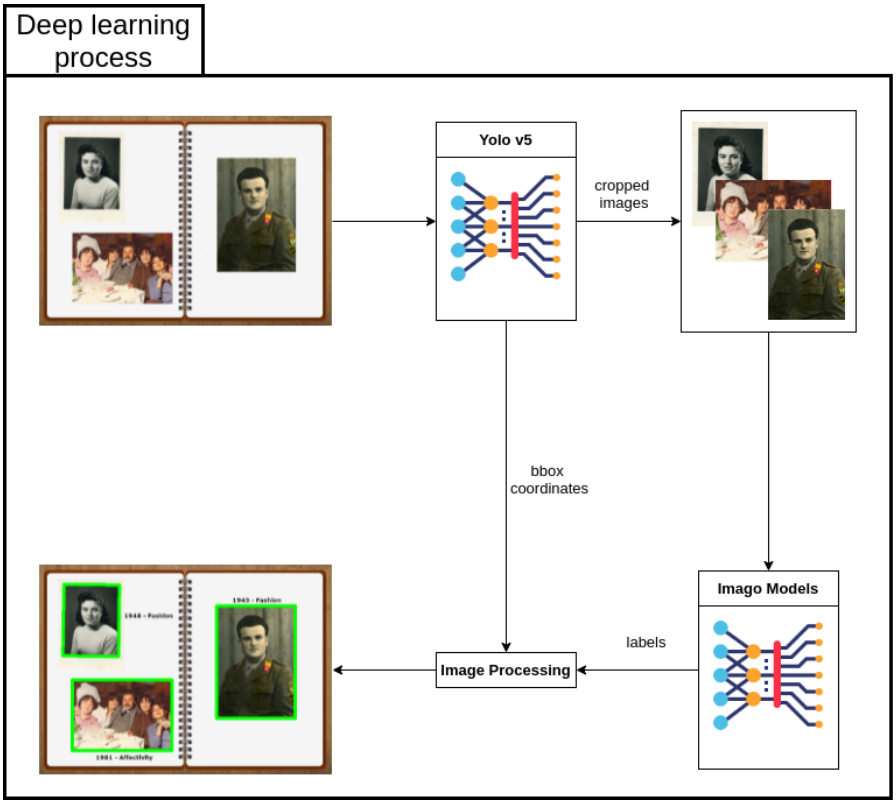


Figure 5.2: Deep learning process architecture.

of deep learning-based object detectors, as they can learn how to manage more varied and complex situations [295]. Within the deep learning object detectors zoo, the YOLO architecture has emerged, since its newest version (v5) [296]. In particular, we resorted to YOLOv5s, because it amounts to a good compromise between performance and memory usage, making it a good candidate to jointly work with the HoloLens 2. The YOLO architecture, however, is not sufficient to solve the task of recognizing photos within family albums, as in the original version it is trained with ImageNet [297]. This motivates the decision of synthesizing a new one, which results from a random pasting, on random backgrounds (e.g., paper, wall, grass backgrounds), of n pictures (with n ranging between 0 and 4), casually picked from the IMAGO dataset. Images might also partially overlap (some examples are reported in Fig. 5.3). With this process, $9,006$ images were obtained and subsequently partitioned in training ($7,372$ images) and test ($1,634$

Figure 5.3: Images from synthetic dataset.

Figure 5.4: Result images from synthetic test set after YOLOv5s inference.

images) sets. Then, we proceeded to fine-tune the YOLOv5s model and exploit data augmentation techniques (e.g., random brightness, horizontal and vertical flipping) for 10 epochs, with a batch size of 32, considering the adam optimizer, and setting a learning rate of 1e-3 with a weight decay equal to 5e-4. A sample of the evaluation of such a trained model on our test set is depicted in Fig. 5.4. The result of this stage is a deep learning model capable of cropping pictures appearing in family albums (top-half of Fig. 5.2).

Photo inference Once the photos were detected, the IMAGO deep learning models, i.e., the IMAGO DATING and IMAGO SOCIO-HISTORICAL CONTEXT classifiers [7] (details in Chapter 4), are exploited to predict the date and

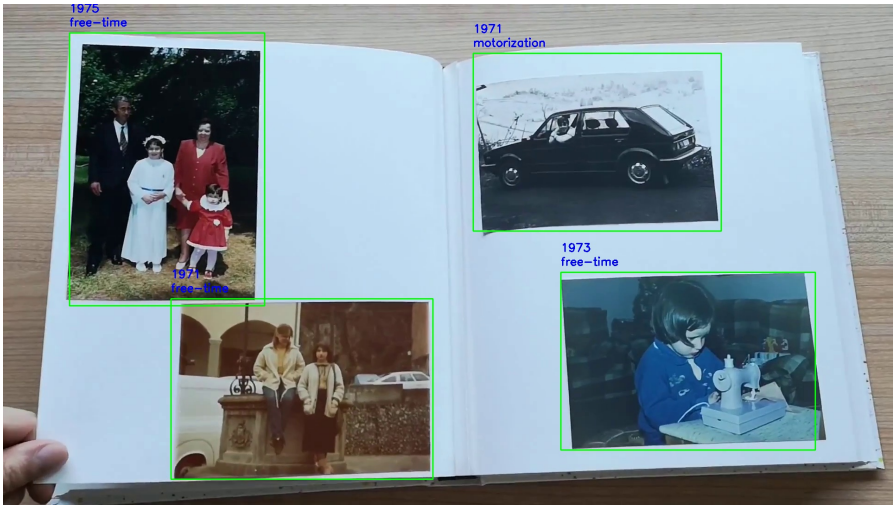


Figure 5.5: Real-world example of augmented HoloLens 2 view.

the socio-historical context of each picture. As specified in Chapter 4, the models are capable of dealing with pictures taken within the 1930-1999 interval, and whose socio-historical context belongs to {*Work*, *Free-time*, *Motorization*, *Music*, *Fashion*, *Affectivity*, *Rites*, *School*, *Politics*}, according to their definition [7] (Section 4.4.2, Chapter 4). Such labels, along with the ones provided by YOLO, are then sent to the HoloLens 2 to augment the view of the photographs with such information (bottom-half of Fig. 5.2).

User-view sharing The labels obtained from the IMAGO deep learning models (i.e., the shooting year and the socio-historical context category) are also leveraged as a piece of information that may be shared, following a collaborative style, and sent to the device interfaces of those users who are viewing photo album from a remote location. To this aim, we built a simple HTTP server to continuously stream, to any kind of device (e.g., smartphone, tablet), the augmented view of the HoloLens 2. In brief, the server processes the video stream captured by the HoloLens 2 and adds to each frame the labels returned by the YOLO and IMAGO deep learning models. The use of HTTP is a design choice meant to support easy access to the stream, from any type of device. A real-world example of the augmented view, as seen from the HoloLens 2, is provided in Fig. 5.5.

5.4.4 CPE - Experimental setting

Participants To evaluate the effectiveness of the proposed AR application, we asked a group of 10 participants to answer some questions regarding their experience. This group had an average age of 26 years, and was composed of 3 females and 7 males. The number of participants has been chosen as a trade-off between the necessity of acquiring sufficient feedback data from a population and the time spent for the evaluation phase. In addition, 10 participants have repeatedly proven to be a sufficient population to discover over 80% of existing interface design problems [298, 299].

Ethics Written consent to participate in this experimental study was collected from each subject. The experimental session was possible thanks to the full compliance with the Covid-19 sanitary protocol adopted by the University of Bologna.

5.4.5 CPE - Experimental evaluation

As aforementioned, once participants tested the experiences, they were asked to complete a survey. This has been designed to assess four constructs: Perceived Ease and Enjoyment of Use (PEEU), Deep Learning Gain (DLG), HoloLens Perspective (HLP), and Receiver Perspective (RP), respectively.

PEEU and DLG constructs The PEEU and DLG constructs were both evaluated through a 5-point Likert scale. For simplicity, a general overview of these constructs is reported in Table 5.1. Individuals' satisfaction and acceptance of a technological innovation, such as an AR application, may be analyzed through different theoretical approaches. The Technology Acceptance Model (TAM) [300] amounts to one of the most popular assessment approaches, as it allows to measure of user intentions in terms of their attitudes, subjective norms, perceived usefulness, perceived ease of use, and related variables. In this project, we want to concentrate on perceived usefulness and ease of use. Perceived usefulness is defined as the degree to which individuals believe that adopting one particular technology will improve an aspect of their life, whereas perceived ease of use is the degree to

Construct	Question	Evaluation
PEEU	(A1) I found the new interface easy to understand	5-point Likert scale
	(A2) I would prefer watching an Augmented Family Photo Album respect to a normal one	5-point Likert scale
	(A3) I enjoyed the overall experience	5-point Likert scale
DLG	(B1) I appreciated the automatic identification of pictures	5-point Likert scale
	(B2) I appreciated the automatic estimate of of pictures' date	5-point Likert scale
	(B3) I appreciated the automatic estimate of of pictures' socio-historical context	5-point Likert scale

Table 5.1: Items and questions used in the survey to assess the Perceived Ease and Enjoyment of Use (PEEU) and Deep Learning Gain (DLG) constructs.

which an individual thinks that adopting a particular technology will be easy to use. Starting from these definitions, we composed the PEEU construct with the following questions:

(A1) I found the new interface easy to understand;

(A2) I would prefer watching an Augmented Family Photo Album with respect to a normal one;

(A3) I enjoyed the overall experience.

The A1 sentence immediately gets to the point, item A2 has been introduced as a further investigation to understand if the users prefer to live an augmented experience with respect to a classical one. Through A3, a broad evaluation of the experience was asked. Following this path, we also want to evaluate the usefulness of the deep learning models that have been developed to carry out the three different computer vision tasks present in this work: family album photos recognition, date, and socio-historical context estimations. For such reason, we also designed the chunk of question items defined as Deep Learning Gain (DLG), which is thought to measure the utility of our deep learning models:

(B1) I appreciated the automatic identification of the pictures;

(B2) I appreciated the automatic estimation of the picture dates;

Construct	Question	Evaluation
HLP	(C1) Would you use the HoloLens 2 application to share your memories?	Yes/No question
	(C2) Nowadays, would you use the HoloLens 2 application to share your photo family album with a distant affection?	Yes/No question
	(C3) Nowadays, would you prefer to share your memories with the HoloLens 2 rather than sharing them in presence?	Yes/No question
	(D1) Would you use the HoloLens 2 application to share with anyone your photo family album?	5-point Likert scale
	(D2) Do you think this HoloLens 2 application would push you to contact more your affections?	5-point Likert scale
	(D3) Do you think this HoloLens 2 application would push you spend more time visualizing your photo family album?	5-point Likert scale
RP	(C4) Would you use this application to revive memories with a distant affection?	Yes/No question
	(C5) Do you think this application would push you to contact more your affections?	Yes/No question
	(D4) Would you use this application to visualize photo family albums of strangers?	5-point Likert scale
	(D5) Nowadays, do you think this application could foster the creation of bonds between strangers?	5-point Likert scale

Table 5.2: Items and questions used in the survey to assess the HoloLens Perspective (HLP) and the Receiver Perspective (RP) constructs.

(B3) I appreciated the automatic estimation of the picture socio-historical context.

HLP and RP constructs The questions regarding both the HLP and RP constructs are defined in Table 5.2. This additional set of questions was defined to explore the different perspectives of users enjoying our application, i.e., the one of the HoloLens 2 wearer and the remote one. In particular, they are based on the concept of Behavioural Intention, which is the individual intention to use a particular technology. Such items are an adaptation of the most significant elements used in [301]. However, different from the previous constructs, which were meant to exclusively measure the usefulness of our system, these questions aim at inspecting more intimate aspects of the users' intentions, i.e., the use they would make of this application and its impact on their daily lives. In particular, both constructs were investigated by exploiting two groups of questions: the C and D groups. The C group is formed by Yes/No question scale questions, to avoid neutral scores:

(C1) Would you use the HoloLens 2 application to share your memories?

(C2) Nowadays, would you use the HoloLens 2 application to share your family photo album with a distant friend or relative?

(C3) Nowadays, would you prefer to share your memories with the HoloLens 2, rather than sharing them in presence?

(C4) Would you use this application to revive memories with a distant affection?

(C5) Do you think this application would push you to contact more your affections?

This group of items appears sufficient to answer and evaluate our constructs. Indeed, they face the problem of sharing memories from different perspectives. Questions C1, C2, and C3 regard the intentions of the HoloLens 2 user. Questions C4 and C5, instead, are about the remote user ones. Nevertheless, we also wanted to explore deeper aspects of Behavioural Intentions. For this reason, we also introduced the D group, evaluated through a 5-point Likert scale, to capture all the nuances of the user's intentions. This set is formed by the following questions:

(D1) Would you use the HoloLens 2 application to share with anyone your family photo album?

(D2) Do you think this HoloLens 2 application would push you to contact more your affections?

(D3) Do you think this HoloLens 2 application would push you to spend more time visualizing your photo family album?

(D4) Would you use this application to visualize family photo albums of strangers?

(D5) Nowadays, may this application help creating bonds between strangers?

The D-items group formed by the first group of questions, D1, D2, and D3, reinforces the opinion regarding the role of our AR application in the revival of the family photo albums cultural phenomena. The second group, instead, composed

Construct	**Items**	$\boldsymbol{\alpha}$	**MIIC**
PEEU	A1-A3	0.73	0.46
DLG	B1-B3	0.81	0.58
HLP	D1-D3	0.69	0.43
RP	D4-D5	0.56	0.39
HLP	C1-C3	0.71	0.45

Table 5.3: Cronbach's α index and MIIC for the considered constructs.

of D4 and D5, regards the possible role that our design could have in socialization, inspecting the possibility to share such intimate material with strangers.

5.4.6 CPE - Experimental results

All the collected data have undergone a reliability check to test their internal consistency and validate our research, through the widely used Cronbach's alpha (α) index. However, α may result in low values for constructs when the tested population is equal to or less than ten items [302]. Therefore, we have also analyzed the Mean Inter-Item Correlation (MIIC), which is appropriate for our case [303]. In a range from 0 to 1, the MIIC confidence interval is 0.15 to 0.50, whereas higher values denote the items overlap. As reported in Table 5.3, all scales demonstrate to be reliable for the MIIC measure (all MIICs > 0.15). As it is possible to see, our analysis does not take into consideration the group C4-C5. This is because such questions are concerning very different aspects. The first one regards the application we are proposing, while the second involves family and personal aspects which are beyond the scope of this research.

PEEU and DLG constructs - survey analysis In Fig. 5.6 are reported the survey results about the PEEU and the DLG construct items. In particular, we have detailed the Mean (M) and the Standard Deviation (SD) for each of them.

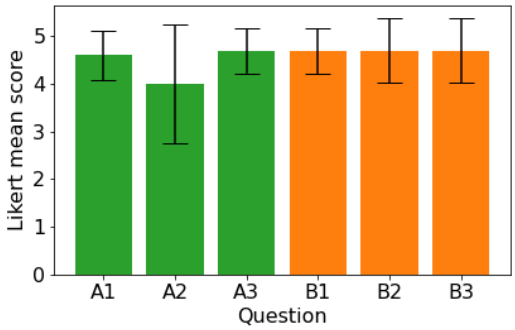
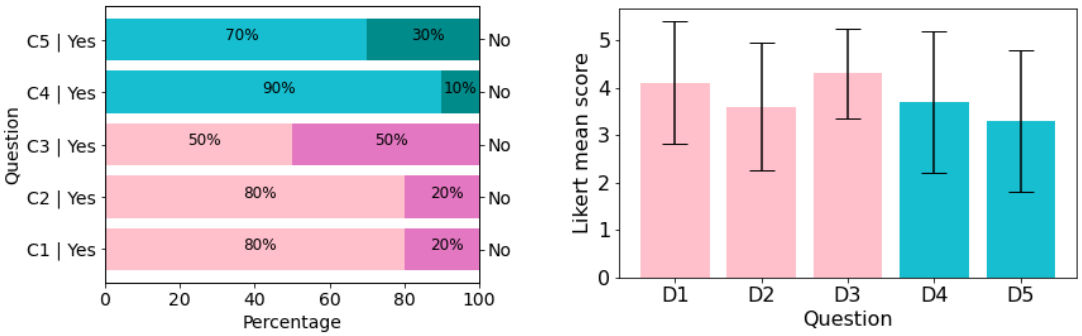


Figure 5.6: Histogram comparison of 5-point Likert A-x and B-x items results, related to the PEEU and DLG constructs, respectively colored in green and orange; the Mean scores, along with their Standard Deviations, are reported.

From such responses, it is evident that there is a strong agreement about the usefulness and ease of use of our application. Indeed, only the A2 item highlights $M < 4$ (where 5 is the maximum). This is because some of the respondents continue to prefer reviving their old memories physically with their affections. Surprisingly, all the questions regarding the DLG construct have $M = 4.5$. This outcome was not so obvious, since the respondents are clearly suggesting their preference for the use of modern technologies in the given application scenario.

HLP and RP constructs - survey analysis Fig. 5.7 reports the survey results for the HLP and the RP construct items. In particular, Fig. 5.7a depicts the percentage of agreement for the C-x items of the two groups, while Fig. 5.7b describes the likelihood for the D-x ones, evaluated with the Mean (M) and the Standard Deviation (SD) of Likert scores. Given the percentage of agreement on the C-x items, reported in Fig. 5.7a, we can infer that the considered population, from both the HoloLens 2 user and the remote perspectives, would use our AR application to contact their affects and revive together their memories, when physically distant. This is of great importance since our work could be useful to bring back to life the tradition of family reunions in front of a family album, even when a family is geographically spread. However, we can notice from the answer to C3, in line with the discussion in the previous analysis, that our respondents were equally



(a) Yes/No answer percentages for the C-x items, related to the HLP and RP constructs, respectively colored in pink and light-blue

(b) Histogram comparison of 5-point Likert D-x items results, related to the HLP and RP constructs, respectively colored in pink and light-blue. The Mean scores, along with their Standard Deviations, are reported

Figure 5.7: HLP and RP construct results.

divided when asked whether they would prefer to live such a moment physically or virtually. The results described in Fig. 5.7b follow the trend of the previous ones. Nevertheless, even if there is great uncertainty (due to high SD), the D2 answer highlights the fact that our proposal may not be sufficiently convincing to contact an affection, in some way linked to the photo album, more than usual. In addition, D4 and D5 scores underline that a large part of our respondents is not so comfortable regarding the sharing of such intimate materials with anyone who wants to appreciate it. Nevertheless, these answers may provide additional inspiration for future works.

5.4.7 Collaborative Photo Environment and digital twins

Starting from this project, from the socio-historical classifiers [7] (photo socio-historical context and shooting year classification) and the Collaborative Photo Environment (CPE) [10], we decided to move in the research direction of Human Collaborative Intelligence (HCLINT) and Digital Twins (DTs). This specific proposal was also published in [11].

The advancements of Artificial Intelligence, Big Data Analytics, and the Internet of Things paved the path to the emergence and use of DTs as technologies to "twin" the life of a physical entity in different fields, both in research and industry [304, 305, 306, 307, 308], aiming at replicating, twinning, or mirroring some physical entity. At the same time, the advent of eXtended Reality (XR) in industrial and consumer electronics has provided novel paradigms that may be put to good use to visualize and interact with DTs. In fact, thanks to XR (i.e., the umbrella term that groups together Virtual, Augmented, and Mixed Reality), it is possible to manipulate DTs directly influencing the physical world and vice versa [305, 306, 308]. Again, XR technologies can support human-to-human interactions for training and remote assistance and could transform DTs into collaborative intelligence tools.

Thinking about our project in such a context, we implemented the Human Collaborative Intelligence empowered Digital Twin framework (HCLINT-DT) integrating human annotations (e.g., textual and vocal) to allow the creation of an all-in-one-place resource to preserve such knowledge. In particular, we concentrated on how humans could help others, adopting the HCLINT-DT approach, since in general, an HCLINT could help humans in supporting other humans in their activities. In fact, *HCLINT involves an extensive collaboration of different team members to solve problems while giving a non-stop real-time learning opportunity*, as reported in [309]. Nevertheless, this framework could be adopted in many fields, not only the cultural heritage one, supporting users to learn how to carry out an unknown process or explore others' past experiences. The assessment of this framework has involved implementing a DT to support human annotations, reflected in both the physical world (AR) and the virtual one (VR). Following this line of thought, we also resorted to a well-known knowledge transfer strategy commonly adopted by humans: asynchronous and persistent annotations. We also moved a further step in this process, providing and sharing human annotations made of text, voice, or videos aligning both the physical and the virtual worlds utilizing XR (AR plus VR, in this case) paradigms. To validate such an approach, we assessed a use case involving family photo albums through an online survey.

As result, the outcomes of the interface design assessment confirm the interest in developing HCLINT-DT-based applications, showing a general agreement on the ease of use of the AR interface and the overall experience, even if there was only a partial agreement in preferring AR. Finally, we also evaluated how the proposed framework could be translated into a manufacturing and industrial context. We explored the adaptability of the proposed approach considering a use case drawn from a local industrial electrical engineering context: the HCLINT-DT showed a good adaptability level.

5.4.8 Discussion and future works

In Section 5.4, we presented an AR system to revive one of the biggest family traditions, i.e., family photo albums exploration, putting to good use the HoloLens 2, with the possibility to exploit a collaborative environment. To reach such a goal, the AR system here proposed includes a trained version of the most known deep learning-based object detector (YOLO, and in particular YOLOv5s) in order to recognize pictures within a family photo album and two additional deep learning models (IMAGO DATING and IMAGO SOCIO-HISTORICAL CONTEXT classifiers [7], already introduced in Chapter 4). Such models served the purpose of providing the information (i.e., date and socio-historical context) needed to augment a given HoloLens 2 user's view, i.e., a family photo album view. In addition, we implemented a simple streaming service, allowing users to access the shared family photo albums from any kind of device (e.g., computer, tablet, and smartphone). The system has been assessed with the interview of ten users who found the interface easy to use and who provided enthusiastic feedback regarding the proposed experience. Based also on the users' comments, we were able to individuate possible future directions of research.

Firstly, as possible future work, we aim to include an active collaboration between HoloLens 2 user and remote ones. In particular, we aim at letting them synchronously manipulate the augmented and shared view, through any kind of non-AR device (e.g., smartphone, tablet, and computer) and AR devices (e.g., HoloLens 2). Such kind of manipulation amounts to provide: (a) data annotation

capabilities through vocal recognition, and (b) affine transformations such as moving, flipping, and rotating. With these extensions, our objective is to increment the level of interest, possibly enhancing the quality of the overall experience. Secondly, we want to augment the capabilities of the examined IMAGO deep learning models, giving them the possibility to infer richer details, such as the people's identity, the country, any symbolic objects (e.g., chairs, cars), and/or specific events (e.g., weddings, birthdays).

Considering, instead, the introduced HCLINT-DT framework (described in Section 5.4.7) [12], possible future research directions include: (a) the integration of crowd intelligence technologies, one of the most promising in the field of Artificial Intelligence and DTs [310, 311], and (b) understanding how the HCLINT-DT framework could support machines (and machine learning), and the consequences of such processes on human activities. Finally, instances of this framework could be created for additional areas and domains, other than the considered ones (e.g., education, marketing). For example, our related work [12] is cited in [312] where the authors aim to provide a holistic approach to design collaboration platforms that may foster multiple communication patterns and workplace productivity to support humans in the loop. In particular, in this case, they focused on the context of the design and management of urban spaces.

5.5 Empowering locksmith crafts through MAR

Considering an artisanal craft rooted back into the past, key locksmithing, a mobile application proposal for key recognition is here presented, showing how Mobile Augmented Reality (MAR) capabilities may today simplify and ameliorate the performances of such ancient trade [12]. In particular, the contents are organized as follows: in Section 5.5.1, the relevant background necessary to explain our MAR application proposal is reported, together with the possible challenges. In Section 5.5.2, we proceed to introduce our proposal, describing the application architecture (Section 5.5.3) and the recognition process (Section 5.5.4). Finally, in Section 5.5.5, we provide an overall discussion to conclude this proposal.

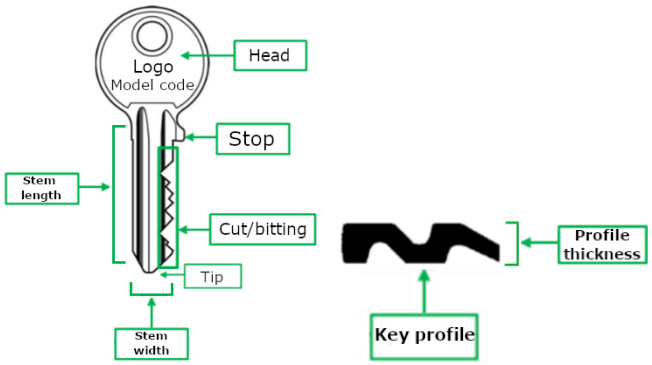


Figure 5.8: Key head, stem and profile main visual features.

5.5.1 Domain knowledge

After providing the background necessary to understand how a key may be selected and recognized, according to specific characteristics, such domain knowledge is translated into technical problems that would emerge adopting MAR and computer vision paradigms. Given a key, the most discriminative parts are the head, stem, and profile. The main characteristics, instead, are reported in Fig. 5.8 and Table 5.4. Then, the key type recognition task consists in identifying the meta-data related to a given key (e.g., serial code, length, silhouette). This task can be simply summarized as follows: (i) take as input picture(s) of a key, and (ii) return as outputs its most probable key type(s) and related meta-data. Nevertheless, many aspects related to this proposal make it challenging: (a) the quality of the pictures could not be sufficient for the intended purpose, (b) distinguishing a key type requires measuring the key stem and profile at a millimeter resolution, and (c) it is mandatory to recognize the silhouettes of different parts of the key (i.e., key head, stem, and profile). In particular, problems (a) and (b) may be addressed considering that the state-of-the-art in mobile devices may today support their solution [313]. Of course, the settings in which the mobile devices will take the key pictures have to be light-safe and homogeneous (e.g., a work table with uniform color). For (c), instead, today it is possible to resort to both computer vision (i.e., contour detection) and machine learning (i.e., K-nearest neighbors) algorithms that may proficiently recognize the silhouettes of all the key parts of interest.

Feature	Description
Serial Code (label)	Unique serial code (alphanumeric)
Head brand	Brand symbol (image)
Head code	Code representing a key profile (alphanumeric)
Head silhouette	Key head silhouette (image)
Stem length	Stem length (measured in mm)
Stem width	Stem width (measured in mm)
Stem silhouette	Key stem silhouette (image)
Profile thickness	Profile thickness (measured in mm)
Profile silhouette	Key profile silhouette (image)

Table 5.4: Key type main discriminative features.

5.5.2 Mobile Key Recognition

Considering the key type recognition task, with this project we want to propose a mobile real-time application, the Mobile Key Recognition (MKR), with a MAR-guided interface, in which an artisan may be assisted step-by-step to reach the key type recognition goal. In the following, the application architecture, and the key recognition process are presented.

5.5.3 MKR - Application architecture

The MAR application here proposed follows a client-server paradigm, where: (i) the client runs on mobile devices (e.g., smartphones, tablets) aiming to capture photos of the key head, stem, and profile; (ii) sends them to the server; (iii) visualizes the returned best matching key type(s). The server side, then: (i) infers the most probable key type(s), taking as input the pictures sent by the client; (ii) sends back to the client MAR interface the inferred key type(s). The MAR application workflow is built based on the design described in Fig. 5.9. Then, it is possible to

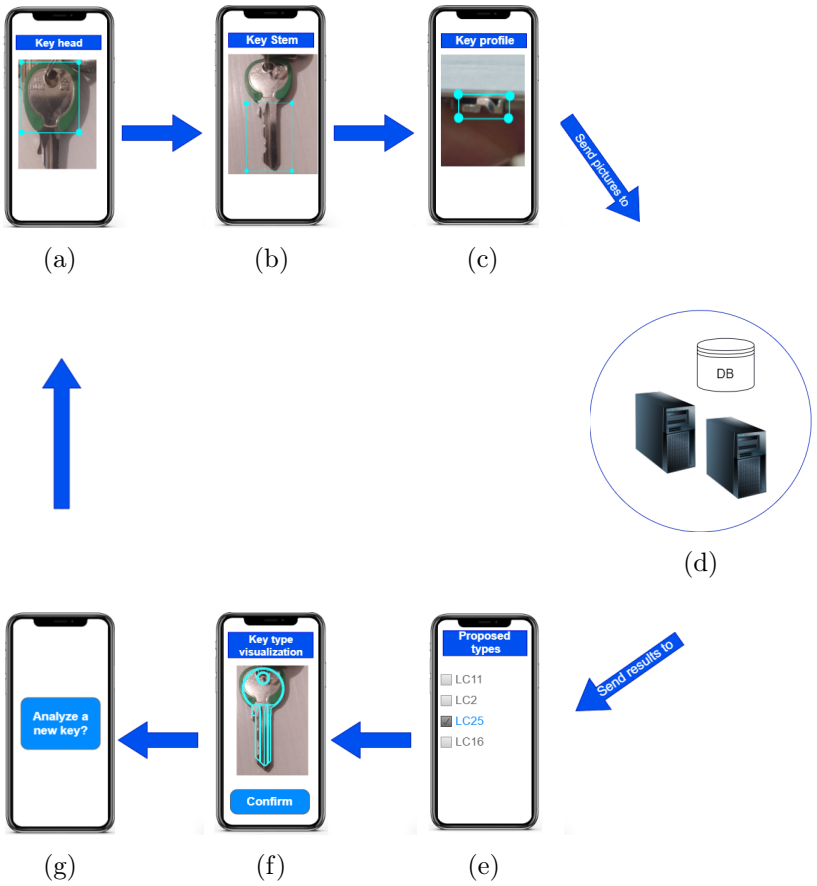


Figure 5.9: Key type recognition MAR application workflow.

distinguish the operations which will contribute to the individuation of the type: the analysis of the key head, stem, and profile, respectively. The information that may be extracted from these key parts is the most discriminative and allow one to search for and recognize the key type (details in Fig. 5.8 and Table 5.4). Such operations are sequential and aim to isolate the most probable key type(s), pruning the options available in the reference database, as more information is exploited at each step. All the previous operations are hidden from the user, who is only asked to point the mobile device aligning a key in the position suggested by the MAR interface. The key alignment is repeated for each region of interest, as depicted in Figs. 5.9a, 5.9b, and 5.9c. The main advantages of a MAR-guided interface are: simplifying the user's work and obtaining the best frame for each analyzed

key part. In particular, the latter is obtained through a segmentation algorithm, following an operation that allows retrieval of the real-size measurements of the key, the camera calibration step. In fact, the camera calibration step, which is performed with the use of a reference object, is necessary to determine the real size of an object in an image. In addition, the reference object, a key in this proposal, should have the following properties: its dimensions have to be known in terms of width or height, and its position should make it easily recognizable (for its position or its intrinsic features like color or shape) [314]. After this calibration step, it is possible to determine all of the key measures (i.e., stem length and width, and profile thickness). Importantly, we remind that such measures fall into a prefixed range, information which is put to good use within the proposed MAR application to support correctly estimate their values. As depicted in Figs. 5.9a, 5.9b, and 5.9c, for each image acquisition step the MAR interface proposes a central shape in which the user is asked to place the different key parts. These shapes are computed by exploiting real-world measurements, thanks to the camera calibration step, and projected with Augmented Reality (AR) techniques. This algorithm aims to isolate the key regions of interest, producing specific pictures which are sent to the server side (Fig. 5.9d). The reference database is organized to facilitate the search operation and, then, it includes only the most discriminative features for the key type recognition task. In particular, it includes two dataframes: the first amounts to a reference one, and it is populated with all the meta-data related to each known key (details in Section 5.5.1), while the second, namely *continual learning dataframe*, is initially empty and it is populated applying a continual learning approach.

5.5.4 MKR - Key recognition process

Following the MAR-guided interface (Figs. 5.9a, 5.9b, and 5.9c), pictures of the key head (both sides), stem, and profile are generated (Fig. 5.10), and after the client side sends them to the server (Fig. 5.9d). In particular, the server iterates until a unique key type is identified or all the pictures are analyzed.

Initially, the pictures regarding the key head are analyzed, as these may reveal

(a) (b) (c)

Figure 5.10: Key (a) head, (b) stem, and (c) profile.

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

Figure 5.11: Key head OCR reader samples.

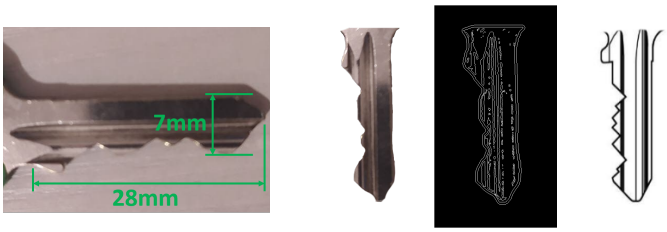
information about the key itself, e.g., brand, text, alphanumeric codes, and even the head silhouette. Such information may dramatically narrow down the set of possible choices for the user. In particular, for what concerns any information printed above the key head, Optical Character Recognition (OCR) modules may be invoked to extract characters or numbers. Some exemplar cases of how an OCR, here EasyOCR [315], may perform are shown in Fig. 5.11. It is possible to note that the OCR module may fail to recognize some characters and/or numbers. For example: in Fig. 5.11a only one character is mistaken; the content of Figs. 5.11b and 5.11c is recognized correctly (excluding vertical text); Fig. 5.11d exhibits only one wrong letter, while no numbers were recognized. Nevertheless, many of these errors could be corrected using algorithms of image processing (e.g., noisy removal, thresholding) and natural language processing (e.g., Levenshtein distance). Then, in essence, the OCR approach appears promising.

(a) (b) (c)

Figure 5.12: Key head (a) picture, (b) silhouette, and (c) actual silhouette.

Successfully, we proceeded to extract the silhouette of the key head. This process is not exclusively related to the key head, but it will also be applied to infer the key stem and profile silhouettes. At this point, pictures of both sides of a key head (an example is given in Fig. 5.12a) are submitted to an image normalization process, which is performed by exploiting a contour detection algorithm [316] that extrapolates an approximation of the silhouette (an example is given in Fig. 5.12b). In this case, the pixels of the contoured images are fed to a machine learning model that exploits pattern similarities between images to retrieve the most probable key type. The aforementioned model was derived training a supervised learning algorithm, namely K-nearest neighbors (KNN) [317], on the silhouettes already included in the reference database (an example is given in Fig. 5.12c).

Then, the choice of a supervised learning approach as the KNN was driven since it is a trade-off between simplicity, low training time, and efficiency [318]. Moreover, the KNN is a suitable algorithm for a continual learning setting. We expect that in many cases, after pruning the reference database with the key head text-like data and silhouette, the search task will end. In this case, the MAR interface shows the best key type. Nevertheless, the approach based on the key head could fail, since only the text-like data and/or the head silhouette is identified, or no useful information could be extracted. If any of these scenarios occur, the next step amounts to analyze the key stem. As described in Section 5.5.1, the key stem is characterized by different parts, but from a purely algorithmic point of view, we could discretize it in three main features: width, length, and silhouette. Thanks to the camera calibration step, the server is already able to measure stem width and



(a) (b1) (b2) (b3)

Figure 5.13: Key stem (a) length and width; key stem (b1) picture, (b2) silhouette, and (b3) actual silhouette.

length. Moreover, the stem silhouette may be extracted through the same process used for the key head. An example of key stem length and width measurements and silhouette extraction is reported in Fig. 5.13. As for the key head analysis, after further pruning the reference database with the key stem information, we hypothesize that the search task could end. In this case, the MAR interface shows the best key type. In case also the key stem analysis fails, the workflow includes the analysis of the key profile, which consists in acquiring its thickness and silhouette. Analogously to the key stem analysis, such information could be guessed with the designed camera calibration, and the silhouette extraction and matching algorithms. An example is reported in Fig. 5.14.

Finally, with this information, the last pruning of the reference database is carried out. No other analyses are possible at this point in the proposed workflow, and full control is taken by the MAR interface. If the key analyses return a unique record, the key type with its meta-data will be displayed on the MAR interface. The possibility of visualizing and manipulating the 3D representation of the predicted type could also be useful to check the correctness of the prediction [319]. Otherwise, if multiple results were provided, the most probable n key types are sorted by importance and then displayed on the MAR interface (Fig. 5.9e). The user can visualize the 3D representation for each of the proposed types, starting from the most probable one (Fig. 5.9f). Therefore, if the first visualized type is not the correct one, the user can choose to visualize the others. For what concerns the handling of wrong answers (none of the most probable key types is the

(a) (b1) (b2) (b3)

Figure 5.14: Key profile (a) thickness; key profile (b1) picture, (b2) silhouette, and (b3) actual silhouette.

correct one), this is a matter of investigation which may be managed depending on the key type. Adopting a continual learning approach the performances of the proposed solution could improve as more data are inserted in the database [320]. The *continual learning dataframe*, used to implement this approach, is composed of rows that contain, for each analyzed key, its corresponding correct type and MAR-extracted images (related to key head, stem, profile, and their respective contours). This dataframe is initially empty and it will be progressively filled by the users that submit new analyses to the MAR system. In particular, this operation will improve the performance of KNN by adding a point belonging to a certain class in the space of the actual data distribution. This should lead to an improvement in cluster robustness, where each cluster represents a key type.

5.5.5 Discussion

In Section 5.5, we introduced a MAR application, the Mobile Key Recognition (MKR) application, to help key locksmiths in their work, recognizing the key types. We integrated the domain into an AR interface, designing an easy-to-use MAR-guided interface. It is worth mentioning that this proposal presents some criticalities. In particular, (a) poor calibration may lead to errors, and (b) the OCR may fail in carrying out its task. Nevertheless, it shows that MAR paradigms may well benefit different kinds of needs and domains impacting the social sphere, also posing the bases of remote locksmithing activities.

5.6 Rethinking wine recognition through AR

Setting in the wine domain, and considering the recognition task, we here present a system to recognize a wine supported by an Augmented Reality (AR) interface [13]. In particular, the contents are organized as follows: in Section 5.6.1 the domain knowledge necessary to explain our project is reported, specifying the information necessary to characterize (and recognize) a wine [13, 14]. In Section 5.6.2 we introduce our framework detailing the design of the AR interface (Section 5.6.3) and all the components, i.e., the machine learning-based Optical Character Recognition (OCR) module, the textual database (textual DB), and the search algorithm (Section 5.6.4). Then, in Section 5.6.5 we show the results obtained with the presented approach, both in terms of efficacy and efficiency. Finally, in Section 5.6.6 we conclude by providing a discussion on how this project may be further improved through possible future works.

5.6.1 Domain knowledge

A wine bottle often includes two labels, a front one and a back one. Typically, the front label is devoted to brand communication, whereas the back label reports all the information characterizing a given wine according to its home-country regulations [321]. In this project, we consider wines bottled following Italian regulations which require specific information to be present on the bottle label (e.g., wine appellation) in the same field of view (i.e., a consumer should not have to turn a bottle to read them all) [322]. Italian wine labels report, in fact, various pieces of information [323, 324, 325, 326]. Follows a list of the most important ones:

- *Name*, typically found at the top-center of the label;

- *Type*, distinguishes wines among wines, varietal wines, and appellation wines. A varietal wine does not possess an appellation and just provides the grape variety used to produce it. In the case of an appellation wine, instead, this is related to the territorial/geographical area of production.

- *Appellation*, can fall into two sub-categories, Protected Geographical Indication (PGI) wines and Protected Designation of Origin (PDO) wines. Italian PDO wines can be DOC or DOCG, now both included in DOP. PGI wines, instead, can be IGT, now included in IGP. Both DOP and IGP indicate products whose characteristics depend on a specific geographical environment, and the second differs from the first because only one of the production phases must take place in that specific area to get this appellation. This information has to be reported on the wine bottle label, but it is possible to choose between DOC, DOCG, or IGT appellations or the corresponding European category, i.e., DOP or IGP. For completeness, we report below the meaning of the following Italian acronyms, translated into English for more clarity: Designation of Origin Controlled (DOC), Designation of Origin Controlled and Guaranteed (DOCG), Protected Designation of Origin (DOP), Typical Geographical Indication (IGT), and Protected Geographical Indication (IGP). In addition, to give a general idea, the *Appellation value* amounts to the "proper name" of the appellation category (e.g., Pignoletto, Romagna). Important to note that on the label the appellation (e.g., DOC, DOP) should appear near the appellation value;

- *Winemaker/winery*, the name of the winery where wine is bottled, should always appear on the label. A winery may work for multiple labels/brands;

- *Region of origin*, is not required. For appellation wines, this information can be inferred from the appellation value. Otherwise, it can be found once the winery has been identified;

- *Origin trademark*, a wine does not necessarily have a trademark of origin. If present, some examples are: Quality Sparkling Wine Produced in a specific region (VSQPRD), Quality Aromatic Sparkling Wine Produced in a specific region (VSAQPRD), Aromatic Sparkling Wine (VSA);

- *Effervescence*, is used to discriminate between still, sparkling, and spumante wines. If nothing is specified on a bottle, the wine is assumed to be still;

- *Sweetness*, is described under different terms based on its type of effervescence. For still and sparkling wines, on one hand, there are terms such as Secco, Semisecco, Abboccato, Amabile, and Dolce. If nothing is specified a still/sparkling wine is considered Secco. For spumante wines, on the other hand, there are many more possible terms, including Brut nature, Extra brut, Brut, Extra dry, Sec, Demi-sec, and Doux. Such information is mandatory only for spumante wines. In addition, if the sugar content of the product justifies the use of two terms, the choice is up to the manufacturer;
- *Color*, can be red, white, or rosé;
- *Mention*, if present, indicates a particular wine characteristic. Some examples are Riserva, Superiore, Classico, and Passito;
- *Bottling year*, is mandatory only for DOP wines;
- *Production method*, if present, this information is often accompanied by the relative logo. Some examples are Organic, Vegan, and No sulfites;
- *Alcohol volume*, is mandatory and expressed as a percentage value. In the possible value range, only .5 steps are allowed;
- *Bottle capacity*, is mandatory. This information and the alcohol volume must be in the same field of view and easily visible (e.g., high color contrast between font and background).

Importantly, the most discriminative information (e.g., wine name, appellation) is usually placed in the top area of the label using a font that is larger than the rest of the text that there appears [323, 324, 325].

5.6.2 Augmented Wine Recognition

The proposed system, the Augmented Wine Recognition (AWR), includes: (a) an AR interface running on a mobile device, and (b) a back-end comprising a hierarchical textual DB and an algorithmic pipeline, that employs a machine learning-based OCR at two different stages, in addition to a specifically customized search

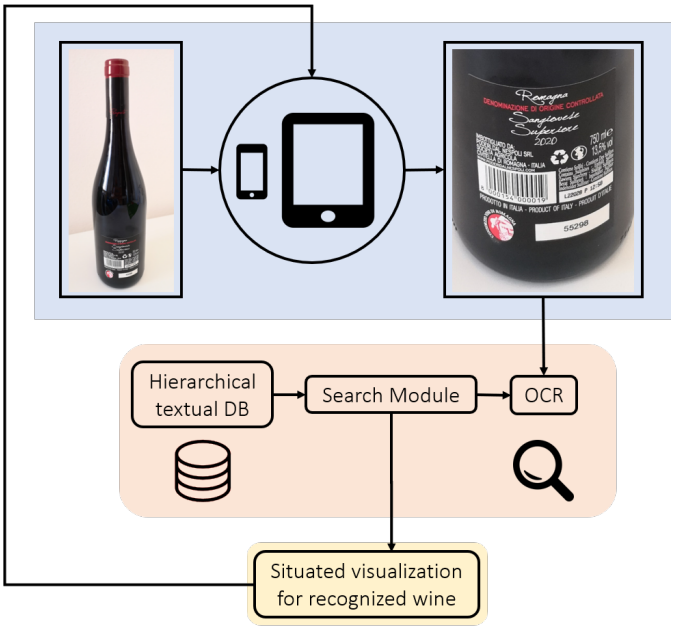


Figure 5.15: The Augmented Wine Recognition (AWR) system.

module. A representative scheme is reported in Fig. 5.15 [13, 14] and, in the following, we present the different components.

5.6.3 AWR - AR interface

The AR interface of the system has been developed for Android-based smartphones. Once started, it continuously scans the surroundings using the mobile camera and, after a certain amount of collected frames, it checks whether the smartphone is targeting a known wine label (showing a spinning loading icon on the bottom left corner of the screen). If a wine label is recognized, the application displays its name, appellation, region, and (if available) the region image associated with the first result of the query. The app also lists other possible candidates on a right panel. In case a different wine result is selected from the list mentioned above, the app displays a dialogue asking to save the selected result and, if confirmed, it shows only the selected wine bottle until the "Close" button is pressed. If the back-end recognizes the targeted wine, but the corresponding entry is not present in the database, an alert is displayed. Lastly, using a toggle on the bottom

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

Figure 5.16: The AR interface: (a) wrong suggestions, (b) correct suggestion, (c) correct wine confirmation, and (d) active scan stops after the correct identification has been confirmed.

right corner of the screen, the user can stop the scan at any moment. In Fig. 5.16 some interface examples are exhibited.

5.6.4 AWR - Back-end

The system back-end comprises a hierarchical textual DB, and an algorithmic pipeline integrating an OCR and a specifically customized search module. In particular, the OCR is involved in two different stages of the pipeline. The first stage, the cropping one, serves the purpose of reducing the area where relevant words will then be searched. This is performed considering that the information useful to individuate a wine is typed with font sizes that are greater than any other text appearing on the label. This fact is then used to create a bounding box that encloses all the pieces of text of interest. The second stage, the decoding one, entails implementing the full pipeline of an OCR, trying to identify the words that

are typed on a label. In this project, we exploited an off-the-shelf deep learning-based OCR, EasyOCR [315], to implement the two steps of interest. In this setting, a hierarchical textual DB can take advantage of the data involved in the process. In fact, the information that characterizes a wine may be organized in mutually exclusive groups and a hierarchical structure, as it emerges from Section 5.6.1. Some pieces of information are mutually exclusive (e.g., *Sweetness*): for example, if a wine is "Dolce" it cannot belong to any of the other categories of sweetness. In addition, if some information appears on a label (e.g., DOC), it is then mandatory to report other pieces of data (e.g., grapes harvest year). Being aware of such types of dependencies and structures, it was possible to drastically reduce the (wine) search space and, consequently, the search running time, as better described in the following. Now, we continue by discussing the components of the AWR back-end, according to the label processing order.

OCR module An OCR algorithm has been employed to initially individuate the spatial locations of relevant words and then the words themselves. The words individuated by the OCR are successively passed to a search algorithm that exploits the textual DB (described in the following). A wealth of research has been performed in the OCR domain [327]. As aforementioned, in this project, we exploited an off-the-shelf deep learning-based OCR, EasyOCR [315], to implement the two steps of interest. The EasyOCR detection component employs the CRAFT algorithm defined in [328], while the recognition model amounts to a CRNN designed in [329], and trained with the pipeline reported in [330]; the decoding step is implemented with CTC [331]. For these reasons, EasyOCR appeared particularly suited to our use case as trained on images belonging to heterogeneous environments (not only scanned documents). In addition, the results reported in [332] identified EasyOCR as the best OCR for natural image scenarios. Then, we proceeded qualitatively to evaluate the potential of this tool on generic wine bottle labels. Notice that the EasyOCR retrieves the text along with the bounding box encapsulating it in the original image. Some qualitative results are reported in Fig. 5.17 [13, 14]. This preliminary experiment showed positive and negative as-

Figure 5.17: EasyOCR retrieved words on different wine labels, accepting all the words, without considering the confidence factor.

pects of EasyOCR. In particular, many relevant keywords were recognized, even with different kinds of fonts and backgrounds, whereas others were not detected or correctly transcribed. This can be motivated in several ways: (a) the text does not lie on a planar surface, (b) the color contrast varies between background and text, (c) poor light conditions, and (d) the words on the wine bottle label could be very distant from the ones used to train the EasyOCR decoder (just some technical words like DOC are included). Hence, EasyOCR may return values that are mistaken and do not hence allow a correct querying mechanism. It may be possible to break such limits by exploiting a fine-tuning approach to the EasyOCR model by resorting to wine domain-specific datasets. However, these datasets should be defined from scratch, labeling wine bottle label pictures with the coordinates of the relevant text and its corresponding characters, resulting in a costly procedure in terms of time and workforce. So, a different path was taken and wrong predictions are corrected by exploiting a search algorithm tailored to pre-defined wine domain dictionary terms. In particular, we will proceed as follow: delineating how the wine database has been organized to expedite the search process and describing the algorithmic pipeline, which also includes the cropping and decoding steps based on EasyOCR.

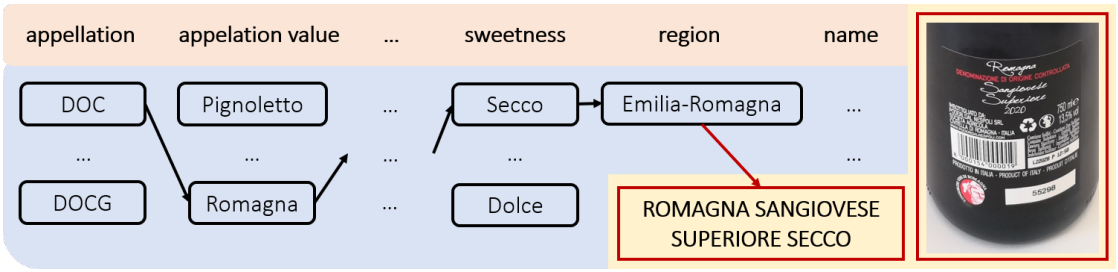


Figure 5.18: Example of a possible wine features conversion from a text-like to a hierarchical-tree-structure.

Wine database The features that distinguish a wine from another are *Type*, *Appellation*, *Appellation value*, *Effervescence*, *Sweetness*, and *Color*. Nevertheless, as anticipated, these features are not independent. For example, the "Lambrusco di Sorbara rosato" is a rosé sparkling wine, with a DOC appellation, but also the "Reggiano Lambrusco rosato" is a rosé sparkling wine, with a DOC appellation. Then, firstly we grouped wines by one feature, like the *Appellation* (e.g., DOC, DOCG), and secondly we created other sub-groups based on each of the other ones, like the *Effervescence* (e.g., sparkling, still, spumante wine). To visually support the previous statements, we report in Fig. 5.18 [13, 14] an example that shows how to convert the wine features from a textual format to a hierarchical-tree-structure. The hierarchical textual DB follows a classical Non-Binary Tree structure. Then, finding a wine entails visiting the last level of the hierarchy. In case of multiple hits (i.e., more than one leaf), these are all returned. Fig. 5.18 shows, in particular, that each layer of the tree is composed of k nodes that represent the possible k values of a particular feature (e.g., the nodes in the level of the *Appellation* represent the values DOC, DOCG, DOP, IGT, and IGP). This choice was driven, not only by the natural match between the wine type hierarchy and tree data structure but also by considering an efficient NoSQL-like database implementation. Our hierarchical tree database follows a specific nested key-value data model, where the key is the value of a feature, and the value is the subset of all the wines that are characterized by that particular value. Knowing a priori the values of the features to traverse, and that the access computational cost in a

one-level key-value database is constant, therefore a single wine-type tree-traversal depends on the number of hierarchy levels (i.e., the features). The hierarchical-tree database can be defined in different ways, according to the order of the features. For this project, the textual DB includes $2,427$ textual descriptions of wine types from the Emilia-Romagna region, built as a four-level hierarchy, considering in order the following features: *Appellation*, *Sweetness*, *Appellation value*, and *Name*. This order matches two criteria: prune as many leaves at the top hierarchy of the tree, and sort features according to their search space. The first one refers to the number of tree nodes pruned when the value of a feature is known. The higher the value, the higher the position of the feature in the hierarchy. The second, instead, refers to the number of possible values that can be matched by the feature. For example, in our dataset, the appellation possesses one among five possible values (i.e., DOC, DOCG, DOP, IGT, and IGP), whereas the appellation value can be one of thirty different values. Hence, the search space of the appellation is lower than the one of the appellation value, and for this reason, it takes a higher place in the hierarchy. These facts motivate the chosen hierarchy, as a trade-off between the tree pruning impact and the matching speed. Finally, to query a particular wine in this dataset, it is sufficient to provide the textual information for the different considered features. At this point, the search algorithm (described in the following) starts finding candidate-relevant words, using a two-step strategy (cropping and decoding) based on EasyOCR. The algorithm continues iterating through the hierarchical textual DB, starting with a linear search that compares such text with the term(s) stored at the given level of the hierarchy, and computing the best match according to a pre-defined textual distance (e.g., Levenshtein, Hamming, Cosine). In practice, each layer of the hierarchy presents a finite set of possible terms and the algorithm picks the most probable one, which depends on the distance between the retrieved words and the terms proposed by the current tree level. Once the best match is found, the algorithm proceeds to analyze the next level of the hierarchy, after pruning the other branches.

(a) Original frame (b) Area of interest (c) Cropped area

Figure 5.19: Example of cropping the area of interest.

Search algorithm Exploiting a domain-specific wine hierarchical textual DB, EasyOCR, and a method to detect and correct OCR post-errors, the full pipeline used to identify a wine is described. As anticipated (details in Section 5.6.1), in a wine bottle label any relevant information is written using a font that is larger than other text and it is usually placed in the top area of the label [325]. Then, for these words, the area retrieved by EasyOCR is larger than any other retrieved one. Based on such a hypothesis we implemented a pre-process step, composed by the *OCR words area detection*, that automatically detects the areas that enclose the words of interest. Not considering this fact would require accounting for all of the words identified on a label (Fig. 5.19a). Instead, with such an assumption it is possible to run the *Cropping by detected areas* algorithm that is responsible for cropping the area of interest. In particular, this is performed as follows. Firstly, only the words enclosed inside bounding boxes that are larger than the median are in the end selected (Fig. 5.19b). Secondly, the bounding box that includes all of such words becomes the final one that is used to crop the label (Fig. 5.19c). In this way, the word detection part has been decoupled from the textual inference one, and the number of words that are processed is

Algorithm 2 Hierarchical search

```
1: procedure WINE_BOTTLE_HIERARCHICAL_SEARCH(wine_bottle_retrieved_words, hierarchical_database,
   wine_bottle_features, threshold, distance_method)
2:     for relevant_feature in wine_bottle_features do
3:         dict_matched_words_values ← {}
4:         for feature_value in relevant_feature.possible_values() do
5:             matched_terms ← LINEAR_SEARCH_POST_OCR_CORRECTION(
                                                    wine_bottle_retrieved_words,
                                                    feature_value, threshold, distance_method)
6:             dict_matched_words_values[feature_value] ← matched_terms
7:         end for
8:         correct_feature_value ← highest_score(dict_matched_words_values)
9:         wine_bottle_retrieved_words, hierarchical_database ← branch_database(
                                                        hierarchical_database,
                                                        wine_bottle_retrieved_words,
                                                        correct_feature_value)
10:    end for
11:    return hierarchical_database
12: end procedure
```

reduced. An initial effect of this choice consists in avoiding words that could be misinterpreted by EasyOCR because of their size and location (e.g., small area words placed near the boundaries of the camera field of view). After this step, the *OCR words inference* is executed, returning all the words found in the images. A successive effect is that the full EasyOCR pipeline is at this point utilized only on a subset of all the words reported on the label, reducing hence the computational cost of this step. To further reduce the set, we remove any duplicate words and some stopwords. Then, the set of words is passed to the final step of the overall algorithm: the *Hierarchical search algorithm*. The python-like pseudo-code for the *Hierarchical search algorithm* is summarized in Algorithm 2. This Algorithm takes as parameters the words returned by the EasyOCR decoder, a copy of the hierarchical textual DB, the wine features, and a distance threshold (Algorithm 2, line 1). With the first cycle, it iterates on the relevant wine features, initializing the dictionary that will then be used to save all matched terms (Algorithm 2, lines 2 and 3). The second cycle (Algorithm 2, line 4), instead, iterates on the possible values of the considered feature (e.g., for the *Appellation*, it will iterate on DOC, DOCG, and IGT). The linear search post-OCR correction algorithm,

described in detail in the following, is then invoked to find any existing matches for the given feature term (Algorithm 2, line 5), and all matches are added to the dictionary (Algorithm 2, line 6). Now, the dictionary contains all the matches found for the given feature values (Algorithm 2, line 7). The algorithm selects the value of the feature that received the highest number of matches. For example, between "Denominazione Origine Controllata" and "Denominazione Origine Controllata e Garantita", the latter would be chosen because more letters have been matched. In case, instead, two features receive the same matching score, they are both selected. The hierarchical database only retains the branch including the matching feature (Algorithm 2, line 8). Once a particular feature value(s) is picked, the hierarchical database is skimmed accordingly, branching the specific sub-tree(s) that involves that value(s). At the end of the skimming procedure, the remaining elements of the hierarchical textual DB contain only one or more elements that possibly include the correct wine. Finally (Algorithm 2, line 9), the hierarchical database is returned. This will contain only one record corresponding to the right bottle of wine or the branches that best match the information found on the label. Given this description, the cost of the *Hierarchical search algorithm* depends on the number of considered features, on all the possible values the considered features may take, and on the cost of the *Linear search post-OCR correction* algorithm that will be discussed shortly. The *Linear search post-OCR correction* algorithm implements two different sub-tasks, detection and correction ones, where the first identifies incorrect tokens, while the second one tries to correct the errors found by the previous step. In this project, this algorithm has been performed by adopting an isolated-word approach, based on a lexical approach relying on specific lexicons (or word unigram language model) and a distance metric for selecting candidates of OCR errors. In this case, the Levenshtein metric has been chosen [333]. The python-like pseudo-code for the *Linear search post-OCR correction* algorithm is reported in Algorithm 3. As also highlighted by this pseudo-code, this algorithm works as a two-level nested loop that iterates over the words retrieved by the OCR and the values of the word that define a wine feature (Algorithm 3, lines 3 and 4). Per each couple, it calculates the distance

Algorithm 3 Linear search post-OCR correction

```
1: procedure LINEAR_SEARCH_POST_OCR_CORRECTION(ocr_retrieved_words, wine_feature_words, threshold, distance_method)
2:     matched_words ← []
3:     for word_wine_feature in wine_feature_words do
4:         for word_ocr in ocr_retrieved_words do
5:             distance ← distance_method(word_ocr, word_wine_feature)
6:             if distance ≤ threshold then
7:                 matched_words.append(word_wine_feature)
8:                 break
9:             end if
10:        end for
11:    end for
12:    return matched_words
13: end procedure
```

adopting the `distance_method` (Algorithm 3, line 5), and if this is less or equal to a certain threshold, the `word_ocr` is considered to be the `word_wine_feature` (Algorithm 3, lines 6 and 7). In this case, the threshold represents a value of accepted distance: the lower the distance, the more likely the `word_ocr` can be considered to be the `word_wine_feature`. In such a case, the algorithm interrupts the inner loop (Algorithm 3, line 8) and continues to search for the next relevant word inside the label text (Algorithm 3, line 3). The threshold value is computed as a percentage of the `word_wine_feature`: a threshold equal to 0 indicates that the two words must be the same (no difference), while a threshold equal to 1 means that `word_ocr` would match `word_wine_feature` as their distance is less or equal than k, where k is the character length of `word_wine_feature`. For what concerns the computational cost, this algorithm processes two lists of respectively n and m size (i.e., `ocr_retrieved_words` and `wine_feature_words`), which are two constants as these may be both upper bounded by some fixed value. Using this search, it is possible to correctly identify terms coming from the wine domain using their eventually distorted form as returned by EasyOCR. Concluding, the *Linear search post-OCR correction* algorithm exploits the domain knowledge (introduced in Section 5.6.1). Some features possess a default value (e.g., the wine *Effervescence*) and, if none of the non-default values is detected, the default one is taken. Also considering this fact, it is worth noticing that failures can be caused

by: (a) a bad identification of a feature (no relevant word is detected in the OCR list or the words are partially wrongly predicted), and (b) relevant information is not reported on the label.

5.6.5 Experimental setting and results

The results obtained exploiting the presented system (described in the previous Sections 5.6.2, 5.6.3, and 5.6.4) on 45 different wine bottles coming from the Emilia-Romagna region in Italy are here reported. The hierarchical textual DB contained $2,426$ different wine coming from this area. The evaluation has been carried out considering a multi-frame setting. In particular, 45 videos are taken, one for each different wine bottle, lasting an average of 8 seconds while rotating the camera around it and selecting one frame per second choosing the less blurred one. This adopts the variance of image Laplacian to get the most in-focus frame, as reported in [334]. In this scenario, the multi-frame setting is motivated by two reasons: (a) employing videos may be possible to reduce OCR detection errors caused by the environment (e.g., light conditions), and (b) it does not require the user to stop and take a picture of what he/she is seeing each time. In the following, the performance of the AWR system is reported both in terms of *efficacy* and *efficiency*.

Efficacy Per each considered frame, the algorithm (described in the previous Sections) is applied to the words retrieved by EasyOCR. This returns all the wines that expose an equal difference between the original name length and the number of words matched (and so the highest number of matches with the words found by the OCR). While calculating the efficacy, a bottle is considered as guessed if its name is included in the set of retrieved ones. Before presenting the obtained results, we must point out that a hyperparameter tuning procedure was carried out to choose the best threshold values that identify a word as recognized (the one introduced in Algorithm 3). To understand what this means in practice, we recall that a word retrieved by EasyOCR is considered to match one in the wine domain set if their Levenshtein distance falls below a given percentage threshold. In the settings of this project, two different thresholds were adopted to recognize

Hyper-parameter	Implemented Model (IM)	IM with a simulated perfect OCR
(Word type, Confidence)	(full words, 0.3), (acronyms, 0.1)	(full words, 0.0), (acronyms, 0.0)
Guessed bottles	41/45	45/45
Accuracy	91%	100%
Returned bottles	min = 1, mean = 1.41, max = 15	min = 1, mean = 1.01, max = 8

Table 5.5: Obtained results with chosen hyperparameters for confidence on the considered wine dataset.

respectively full words and acronyms (e.g., Denominazione di Origine Controllata vs. DOC). Hence, the hyperparameter tuning consists in testing the algorithm on all the considered bottles, varying those confidences between 0 and 1, with a step of .1. The confidences picked are respectively 0.3 for the full words and 0.1 for the acronym, which returned the best overall results: 41/45 bottles were guessed, reaching a 91% of accuracy. This accuracy value was obtained, considering that the algorithm could provide more than one bottle as output, retrieving on average 1.41 bottles per examined video (with a maximum of 15 and a minimum of 1). It is interesting to note that the errors depended on the performance of the OCR. To prove this, we created 45 different text files, one per tested bottle, to mimic the performance of a perfect OCR, simply transcribing all the words contained in the bottle respecting the order of the considered language (i.e., Italian). Adopting such data and imposing the confidence to 0 (with a perfect OCR setting the algorithm should match the exact words), the algorithm did not provide any error. On average, the linear search algorithm applied to the perfect OCR setting retrieved on average 1.01 bottles per examined video (with a maximum of 8 and a minimum of 1). All the presented results are reported in Table 5.5.

Efficiency In such a context, it is important to measure not only the efficacy but also the efficiency of the considered algorithm. As stated in the previous Sections, the algorithm is composed of four steps: (i) EasyOCR words area detection, (ii) cropping by detected areas, (iii) EasyOCR words inference, and (iv) final hi-

Mode	Step	Min time [s]	Mean time [s]	Max time [s]
Hierarchical	OCR words area detection	0.44	0.55	0.69
	Cropping by detected areas	0.025	0.029	0.036
	OCR words inference	0.45	0.83	1.24
	Hierarchical search algorithm	0.11	0.96	2.15
	Total time	1.02	2.37	4.11
Full-linear	**Total time**	22.63	153.66	404.53

Table 5.6: Min, mean and max time of the bottle detection algorithm over the considered bottles and frames, reported in s.

erarchical search algorithm. For this reason, we here report the mean time taken by the entire algorithm pipeline to identify an examined bottle. Importantly, all the captured times regarding EasyOCR do not include the deep learning model initialization times. Time measurements were obtained by executing a single trial of the complete algorithmic pipeline on all the bottles video frames. In Table 5.6 the min, mean and max of such times, calculated over the bottles, are reported in seconds (grouped under the *Hierarchical* mode). Each metric derives from the average time over 10 trials, and this process was done to better approximate the effective distribution of such times. At this point, a comparison between the efficiency of the hierarchical tree search vs. a classical linear one has been performed. In particular, we defined a single-column database that contains all the relevant values of the different features that discriminate a wine. Instead of pruning the tree based on given feature values, we applied Algorithm 2 at each step of the whole data structure. Then, the final output consists of all those wines that have matched the maximum number of their word description. Again, we report the computed min, mean, and max values obtained over the same set used for previous evaluations over 10 trials. The obtained results are included at the bottom of Table 5.6 (under the *Full-linear* mode). These times are roughly a hundred times

higher than the ones presented before. This is justified by the fact that the tree is never pruned, and consequently, the algorithm compared all the OCR retrieved words with all the m words that compose each of the n samples ($2,426$ in our case).

5.6.6 Discussion and future works

In Section 5.6, we presented the Augmented Wine Recognition (AWR) system, and all the details regarding both the domain knowledge (e.g., wine background) and the technical one (e.g., OCR approach) are provided. In addition, we reported various related works to underline the aspects in which our system differs from others [13, 14]. Nowadays (augmented) wine recognition systems and services are typically based on pure computer vision approaches, but not without limitations. Such limitations, hence, have motivated this project. In particular, a wine could possess different visual features (in time), but many wines could appear similarly. A sole visual analysis may cause wrong detections, because the labels change (not a rare phenomenon), also due to the changes in the related regulations. To avoid this phenomenon, very frequent database/fine-tuning updates would be required. This may not always be possible, though, as in the case of rare labels belonging to the long tail of the market. Adopting a textual perspective, hence, allowed us to overcome limitations posed by visual features, expanding the possible field of application, both in time and space, and speeding up the update process. The results appear promising, but more work is needed to improve and generalize our system in a more varied context.

Firstly, further works on image pre-processing techniques may be carried out to improve the Optical Character Recognition (OCR) performance (e.g., image rectification [335] and super resolution [336]). This may also involve the adoption of deep learning models for the semantic segmentation of wine labels, following an approach similar to [337]. Following this point, a larger dataset of labeled bottle images, which also includes the annotations of the words along with (specifying the language) their bounding boxes, should be created. Adopting such a dataset may ease several problems. The OCR performance could increase by exploiting a fine-tuning approach using label visual annotations and domain-specific vocabularies.

This would ameliorate the problem highlighted in the results reported in Table 5.5. This will also pose the bases for a more general and applicable detection pipeline that exceeds the Italian wines domain (e.g., fine-tuning in other languages).

Secondly, it may be possible to adopt a multi-domain approach for wine recognition exploiting our OCR-based framework, but also deep learning models to analyze visual features similarity (i.e., image retrieval). One of the possible use cases of this approach can be described by the following steps: (i) to use the presented textual framework to skim, as much as possible, the textual database (textual DB), and (ii) to execute a simple query image on the remaining wines in another image database.

Finally, by correctly adopting the OCR, other relevant information in the label can be detected, e.g., the *Bottle capacity* and the *Alcohol content*. These could be used, along with the wine *Name*, to derive other information such as the calorie intake, maximum recommended daily dose, and maximum dose for driving. Hence, many possible future research directions could be explored to improve the proposed system. Nevertheless, promising performance is achieved showing that an identification mechanism based not only on visual features but also on textual data could be a valid method in many application contexts.

5.7 Conclusions

In these projects, we have proposed (three) different AR/MAR applications and systems in (three) different contexts, starting from a Collaborative Photo Environment [10, 11] in the cultural heritage field, proceeding with a Mobile Key Recognition [12] application considering an artisan craft, i.e., the key locksmithing, and ending with an Augmented Wine Recognition [13] system setting in the wine domain. The principal aim of all these projects was to prove (and highlight) the benefits in different terms (e.g., efficacy, efficiency, performance, and usability), principally combining machine learning algorithms and eXtended Reality interfaces, regardless of the application context.

Chapter 6

Conclusions

Starting from what was stated in the Introduction (Chapter 1), we here draw the conclusion of this Thesis research work. Since each previous Chapter contains one (or more) detailed Section(s) related to the Discussion, Conclusions, and Future works, the following are reported general conclusions concerning the overall research work, in order to avoid unnecessary repetitions. The main focus was to prove that defining machine learning as the class of learning algorithms [17] is limiting, and not enough. In particular, considering the principal elements characterizing a machine learning pipeline, from the developers to the users, from the task to the performance measure, we aimed at analysing machine learning in its entirety. Importantly, we followed this way continuously searching for connections with real and pragmatic applications, from different perspectives and points of view.

From a non-data scientist point of view, and a developer perspective, *considering a Human Activity Recognition task in the Sport Science area*, the complexity of applying machine (and deep) learning algorithms is shifting more and more from possessing a data science knowledge and algorithm implementation background to owning the ability to design systems and collect data (Chapter 2). This is due to the increasing development of machine learning tools and platforms: Weka, Orange, Ludwig, and Knime, for example. The analysis showed how the employed machine learning algorithms with a raw dataset can be powerful but, simultaneously, parameter-dependent. Data-human interfaces combined with machine

learning algorithms may certainly represent useful instruments for non-data scientists nevertheless, they should be utilized with attention. Further investigations could be done, considering additional datasets, platforms, and algorithms.

From the task perspective, and to structuring (and organizing) a new one from scratch, the possibility to utilize a low-cost wearable to collect raw data, *considering a Human Reaching Movement analysis in the Psychological field*, to set a complete stage for a learning task (waiting to reach a sufficient amount of data), was analyzed in detail (Chapter 3). The obtained results support the use of a commercial and low-cost 3-axis accelerometer to collect raw data in the considered task, posing the basis for similar studies in similar settings. Instead, from a performance perspective, future studies would benefit from the use of additional kinematic indices, with the possibility to include video recordings and offline coding to capture additional dimensions and data. Following this line, building a task to apply machine learning is where "we arrive" from this research (now), but it is also from where "we start". The obtained results want to be the initial step to exploring through machine learning specific aspects of human reaching movement, e.g., the distinctive contribution of motor planning and control. Machine learning algorithms could study (and learn information) from different input data. In this case, a supervised data set would incrementally improve the results, but also an unsupervised approach could be taken into consideration.

Starting from a non-data scientist developer perspective, through setting a task from scratch, we arrived to consider the model point of view, including the domain knowledge, the architecture, the performance measure, and all the relevant elements that characterize a learning pipeline. In particular, we explored this "point of view" *considering Socio-Historical aspects in the Cultural Heritage environment* (Chapter 4). We proposed a multimedia application to assist socio-historians in cataloging family album photos. We trained and tested single and multi-input (ensemble and merged) deep learning models, considering both convolutional neural network and transformer-based classifiers. The obtained results proved that our application could provide important support to socio-historians. In addition, we concentrated on how quantitative methods (deep learning-based classifiers) may

support qualitative analysis (socio-historical aspects), and the possibility to search for cues of intercultural influences through cross-dataset experiments, obtaining in both cases interesting results. This only represents a step in the direction of creating a holistic approach to socio-historical tasks and, analyzing the limits of this research, further investigations in this domain may consider larger amounts and more balanced sets of data, better segmentation of the relevant areas of the images, and the implementation of a multi-modal approach, capable of including also other sources of information and data formats, e.g., text.

Finally, we explored the point of view of users, and how machine learning can improve the interface user experience through eXtended Reality. In particular, we proceeded *considering Augmented Reality, and Mobile Augmented Reality interface applications in different contexts*: the cultural heritage and family album photos, the artisans' crafts with the key locksmith figure, and the wine domain through bottle labels (Chapter 5). We aimed at proving and highlighting the benefits to combine machine (and deep) learning algorithms and eXtended Reality-guided interfaces, regardless of the application context. The eXtended Reality, in fact, is transforming into a technology that may be available in a variety of contexts, expanding from an only academic or highly specialized technology to an everyday one.

Then, we proved that talking about machine learning, we do not necessarily want to focus on an algorithm or an application for its own sake as the goal of the discussion. In addition, not only in a quantitative task, not only in a scientific environment, and not only from a data scientist perspective, machine (and deep) learning can do a difference. From this, it is possible to proceed in different directions to search other machine learning perspectives and points of view, as well as other application contexts. This is not said to be the end, it could be a new beginning.

Acknowledgements

This PhD research work was supported by the University of Bologna and the Golinelli Foundation of Bologna.

Bibliography

[1] N. Riedel, A. Angeli, and G. Marfia, "Qualitative activity recognition using machine and deep learning: Experimenting with data-human interfaces for non data-scientists," in *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 7–12, 2019.

[2] A. Angeli, G. Marfia, and N. Riedel, "Using off-the-shelf data-human interface platforms: traps and tricks," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 12907–12929, 2021.

[3] A. Angeli, I. Valori, T. Farroni, and G. Marfia, "Reaching to inhibit a prepotent response: A wearable 3-axis accelerometer kinematic analysis," *Plos one*, vol. 16, no. 7, p. e0254514, 2021.

[4] I. Valori, A. Angeli, T. Farroni, and G. Marfia, "Kinematics of reaching movements in children with adhd: motor planning and control," in *Budapest CEU Conference on Cognitive Developement, Poster presentation*, 2021.

[5] I. Valori, A. Angeli, T. Farroni, and G. Marfia, "Do i need reward or control? action selection, motor planning and control in children with adhd," in *Budapest CEU Conference on Cognitive Developement, Poster presentation*, 2022.

[6] I. Valori, L. Della Longa, A. Angeli, G. Marfia, and T. Farroni, "Reduced motor planning underlying inhibition of prepotent responses in children with adhd," *Sci rep*, vol. 12, no. 18202, 2022.

[7] L. Stacchio, A. Angeli, G. Lisanti, D. Calanca, and G. Marfia, "Towards a holistic approach to the socio-historical analysis of vernacular photos," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.

[8] L. Stacchio, A. Angeli, G. Lisanti, and G. Marfia, "Applying deep learning approaches to mixed quantitative-qualitative analyses," in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pp. 161–166, 2022.

[9] L. Stacchio, A. Angeli, G. Lisanti, and G. Marfia, "Searching for cultural relationships through deep learning models," in *Proceeding of the 1st International Virtual Conference on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding*, `http://ceur-ws.org/Vol-3266/`, 2022.

[10] L. Stacchio, A. Angeli, S. Hajahmadi, and G. Marfia, "Revive family photo albums through a collaborative environment exploiting the hololens 2," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 378–383, IEEE, 2021.

[11] L. Stacchio, A. Angeli, and G. Marfia, "Empowering digital twins with extended reality collaborations," *Virtual Reality & Intelligent Hardware*, vol. 4, no. 6, pp. 487–505, 2022.

[12] L. Stacchio, A. Angeli, and G. Marfia, "Empowering locksmith crafts via mobile augmented reality," in *Proceedings of the Conference on Information Technology for Social Good*, pp. 305–308, 2021.

[13] L. Stacchio, A. Angeli, L. Donatiello, A. Giacche, and G. Marfia, "Rethinking augmented wine recognition," in *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 560–565, IEEE, 2022.

[14] A. Angeli, L. Stacchio, L. Donatiello, A. Giacchè, and G. Marfia, "Making

paper labels smart for augmented wine recognition," *The Visual Computer*, 2023, *submitted*.

[15] E. Velloso, A. Bulling, H. Gellersen, W. Ugulino, and H. Fuks, "Qualitative activity recognition of weight lifting exercises," in *Proceedings of the 4th Augmented Human International Conference*, pp. 116–123, 2013.

[16] "Geneactiv." `https://www.activinsights.com/products/geneactiv/`.

[17] T. Mitchell, *Machine Learning*. McGraw-Hill International Editions, McGraw-Hill, 1997.

[18] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[20] M. García, C. Domínguez, J. Heras, E. Mata, and V. Pascual, "An on-going framework for easily experimenting with deep learning models for bioimaging analysis," in *International Symposium on Distributed Computing and Artificial Intelligence*, pp. 330–333, Springer, 2018.

[21] A. Naik and L. Samant, "Correlation review of classification algorithm using data mining tool: Weka, rapidminer, tanagra, orange and knime," *Procedia Computer Science*, vol. 85, pp. 662–668, 2016.

[22] I. Guyon, I. Chaabane, H. J. Escalante, S. Escalera, D. Jajetic, J. R. Lloyd, N. Macià, B. Ray, L. Romaszko, M. Sebag, *et al.*, "A brief review of the chalearn automl challenge: any-time any-dataset learning without human intervention," in *Workshop on Automatic Machine Learning*, pp. 21–30, 2016.

[23] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: a survey of case studies," *ACM Computing Surveys (CSUR)*, 2020.

[24] C. Crisci, B. Ghattas, and G. Perera, "A review of supervised machine learning algorithms and their applications to ecological data," *Ecological Modelling*, vol. 240, pp. 113–122, 2012.

[25] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Computer methods and programs in biomedicine*, vol. 161, pp. 1–13, 2018.

[26] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017.

[27] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[28] L. Zhu, P. Spachos, E. Pensini, and K. N. Plataniotis, "Deep learning and machine vision for food processing: A survey," *Current Research in Food Science*, vol. 4, pp. 233–249, 2021.

[29] G. D. Styliaras, "Augmented reality in food promotion and analysis: Review and potentials," *Digital*, vol. 1, no. 4, pp. 216–240, 2021.

[30] K. Nuelle, S. Bringeland, S. Tappe, B. Deml, and T. Ortmaier, "Mobile augmented reality system for craftsmen," in *Developing Support Technologies*, pp. 169–176, Springer, 2018.

[31] N. Rohrbach, P. Gulde, A. R. Armstrong, L. Hartig, A. Abdelrazeq, S. Schröder, J. Neuse, T. Grimmer, J. Diehl-Schmid, and J. Hermsdörfer, "An augmented reality approach for adl support in alzheimer's disease: a crossover trial," *Journal of neuroengineering and rehabilitation*, vol. 16, no. 1, pp. 1–11, 2019.

[32] M. Zaher, D. Greenwood, and M. Marzouk, "Mobile augmented reality applications for construction projects," *Construction Innovation*, vol. 18, pp. 152–166, 01 2018.

[33] V. Paelke, "Augmented reality in the smart factory: Supporting workers in an industry 4.0. environment," in *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, pp. 1–4, 2014.

[34] A. Bujari, B. Licar, and C. E. Palazzi, "Movement pattern recognition through smartphone's accelerometer," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 502–506, IEEE, 2012.

[35] A. Bujari, B. Licar, and C. E. Palazzi, "Road crossing recognition through smartphone's accelerometer," in *2011 IFIP Wireless Days (WD)*, pp. 1–3, IEEE, 2011.

[36] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013.

[37] W. M. Van der Aalst, "Data scientist: The engineer of the future," in *Enterprise interoperability VI*, pp. 13–26, Springer, 2014.

[38] M. Zorrilla and D. García-Saiz, "A service oriented architecture to provide data mining services for non-expert data miners," *Decision Support Systems*, vol. 55, no. 1, pp. 399–411, 2013.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[40] N. Ketkar, "Introduction to pytorch," in *Deep learning with python*, pp. 195–208, Springer, 2017.

[41] N. Ketkar, "Introduction to keras," in *Deep Learning with Python*, pp. 97–111, Springer, 2017.

[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[43] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[44] G. Marfia and M. Roccetti, "A practical computer based vision system for posture and movement sensing in occupational medicine," *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 8109–8129, 2017.

[45] M. Kroes, A. G. Kessels, A. C. Kalff, F. J. Feron, Y. L. Vissers, J. Jolles, and J. S. Vles, "Quality of movement as predictor of adhd: results from a prospective population study in 5-and 6-year-old children," *Developmental Medicine and Child Neurology*, vol. 44, no. 11, pp. 753–760, 2002.

[46] G. Buscher, S. T. Dumais, and E. Cutrell, "The good, the bad, and the random: an eye-tracking study of ad quality in web search," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49, ACM, 2010.

[47] W. Ugulino, E. Velloso, and H. Fuks, "Human activity recognition," 2019.

[48] "Weka." `https://www.cs.waikato.ac.nz/ml/weka/`, 1993.

[49] "Orange." `https://orange.biolab.si/`, 1996.

[50] "Ludwig Deep Learning." `https://uber.github.io/ludwig/`, 2019.

[51] "KNIME Open for Innovation." `www.knime.com`, 2006.

[52] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[53] J. Heaton, N. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3–12, 2017.

[54] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS computational biology*, vol. 3, no. 6, p. e116, 2007.

[55] S. Athey, "The impact of machine learning on economics," in *The economics of artificial intelligence: An agenda*, University of Chicago Press, 2018.

[56] M. Bohanec, M. K. Borštnar, and M. Robnik-Šikonja, "Explaining machine learning models in sales predictions," *Expert Systems with Applications*, vol. 71, pp. 416–428, 2017.

[57] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[58] K. Patel, "Lowering the barrier to applying machine learning," in *Adjunct proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 355–358, ACM, 2010.

[59] Q. Yang, J. Suh, N.-C. Chen, and G. Ramos, "Grounding interactive machine learning tool design in how non-experts actually build models," in *Proceedings of the 2018 on Designing Interactive Systems Conference 2018*, pp. 573–584, ACM, 2018.

[60] D. Chen, R. K. Bellamy, P. K. Malkin, and T. Erickson, "Diagnostic visualization for non-expert machine learning practitioners: A design study," in *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 87–95, IEEE, 2016.

[61] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer learning for activity recognition: A survey," *Knowledge and information systems*, vol. 36, no. 3, pp. 537–556, 2013.

[62] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert systems with applications*, vol. 59, pp. 235–244, 2016.

[63] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[64] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1488–1492, IEEE, 2015.

[65] N. D. Lane, P. Georgiev, and L. Qendro, "Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 283–294, ACM, 2015.

[66] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[67] T. Plötz, N. Y. Hammerla, and P. L. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[68] N. Y. Hammerla, J. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plötz, "Pd disease state assessment in naturalistic environments using deep learning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[69] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450–457, 2014.

[70] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[71] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, pp. 1–17, Springer, 2004.

[72] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Aaai*, vol. 5, pp. 1541–1546, 2005.

[73] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille, "A long-term evaluation of sensing modalities for activity recognition," in *International conference on Ubiquitous computing*, pp. 483–500, Springer, 2007.

[74] B. Tessendorf, F. Gravenhorst, B. Arnrich, and G. Tröster, "An imu-based sensor network to continuously monitor rowing technique on the water," in *2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 253–258, IEEE, 2011.

[75] M. Kranz, A. MöLler, N. Hammerla, S. Diewald, T. PlöTz, P. Olivier, and L. Roalter, "The mobile fitness coach: Towards individualized skill assessment using personalized mobile devices," *Pervasive and Mobile Computing*, vol. 9, no. 2, pp. 203–215, 2013.

[76] C. Ladha, N. Y. Hammerla, P. Olivier, and T. Plötz, "Climbax: skill assessment for climbing enthusiasts," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 235–244, ACM, 2013.

[77] S. L. Lau, I. König, K. David, B. Parandian, C. Carius-Düssel, and M. Schultz, "Supporting patient monitoring using activity recognition with

a smartphone," in *2010 7th International Symposium on Wireless Communication Systems*, pp. 810–814, IEEE, 2010.

[78] M. Sung, C. Marci, and A. Pentland, "Wearable feedback systems for rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 2, no. 1, p. 17, 2005.

[79] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ecg features for screening paroxysmal atrial fibrillation patients," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2095–2104, 2017.

[80] I. Maurtua, P. T. Kirisci, T. Stiefmeier, M. L. Sbodio, and H. Witt, "A wearable computing prototype for supporting training activities in automotive production," in *4th International Forum on Applied Wearable Computing 2007*, pp. 1–12, VDE, 2007.

[81] T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing*, no. 2, pp. 42–50, 2008.

[82] A. Khan, S. Mellor, E. Berlin, R. Thompson, R. McNaney, P. Olivier, and T. Plötz, "Beyond activity recognition: skill assessment from accelerometer data," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1155–1166, ACM, 2015.

[83] "Python." `https://www.python.org/`.

[84] J. Ortiz Laguna, A. G. Olaya, and D. Borrajo, "A dynamic sliding window approach for activity recognition," in *User Modeling, Adaption and Personalization* (J. A. Konstan, R. Conejo, J. L. Marzo, and N. Oliver, eds.), (Berlin, Heidelberg), pp. 219–230, Springer Berlin Heidelberg, 2011.

[85] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A machine learning workbench," in *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference*, pp. 357–361, IEEE, 1994.

[86] J. Demšar, B. Zupan, G. Leban, and T. Curk, "Orange: From experimental machine learning to interactive data mining," in *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 537–539, Springer, 2004.

[87] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, *et al.*, "Orange: data mining toolbox in python," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2349–2353, 2013.

[88] P. Molino, Y. Dudin, and S. S. Miryala, "Ludwig Deep Learning," 2019.

[89] A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum, and M. R. Berthold, "Knime for reproducible cross-domain analysis of life science data," *Journal of biotechnology*, vol. 261, pp. 149–156, 2017.

[90] G. Rossum, "Python reference manual," 1995.

[91] "Scikit-learn." `www.scikit-learn.org`, 2007.

[92] "Keras." `www.keras.io`, 2015.

[93] F. Malik, "Neural networks: a solid practical guide," 2019.

[94] S. Glover, "Separate visual representations in the planning and control of action," *Behavioral and brain sciences*, vol. 27, no. 1, pp. 3–24, 2004.

[95] P. Cisek, "Cortical mechanisms of action selection: the affordance competition hypothesis," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 362, no. 1485, pp. 1585–1599, 2007.

[96] K. R. Ridderinkhof, W. P. Van Den Wildenberg, S. J. Segalowitz, and C. S. Carter, "Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning," *Brain and cognition*, vol. 56, no. 2, pp. 129–140, 2004.

[97] N. P. Friedman and A. Miyake, "The relations among inhibition and interference control functions: a latent-variable analysis.," *Journal of experimental psychology: General*, vol. 133, no. 1, p. 101, 2004.

[98] L. F. Koziol, D. E. Budding, and D. Chidekel, "From movement to thought: executive function, embodied cognition, and the cerebellum," *The Cerebellum*, vol. 11, no. 2, pp. 505–525, 2012.

[99] E. Thelen, "Time-scale dynamics and the development of an embodied cognition," *Mind as motion: Explorations in the dynamics of cognition*, pp. 69–100, 1995.

[100] P. Wilson, S. Ruddock, S. Rahimi-Golkhandan, J. Piek, D. Sugden, D. Green, and B. Steenbergen, "Cognitive and motor function in developmental coordination disorder," *Developmental Medicine & Child Neurology*, vol. 62, no. 11, pp. 1317–1323, 2020.

[101] E. M. Sokhadze, A. Tasman, G. E. Sokhadze, A. S. El-Baz, and M. F. Casanova, "Behavioral, cognitive, and motor preparation deficits in a visual cued spatial attention task in autism spectrum disorder," *Applied psychophysiology and biofeedback*, vol. 41, no. 1, pp. 81–92, 2016.

[102] A. Diamond, "Executive functions," *Annual review of psychology*, vol. 64, p. 135, 2013.

[103] G. Leisman, A. A. Moustafa, and T. Shafir, "Thinking, walking, talking: integratory motor and cognitive brain function," *Frontiers in public health*, p. 94, 2016.

[104] R. A. Barkley, "Response inhibition in attention-deficit hyperactivity disorder," *Mental retardation and developmental disabilities research reviews*, vol. 5, no. 3, pp. 177–184, 1999.

[105] A. P. Association *et al.*, "American psychiatric association: Diagnostic and statistical manual of mental disorders, arlington," 2013.

[106] A. E. Doyle, "Executive functions in attention-deficit/hyperactivity disorder," *Journal of Clinical Psychiatry*, vol. 67, p. 21, 2006.

[107] J. Nigg, "On the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology," *Journal of Child Psychology and Psychiatry*, vol. 58, no. 4, pp. 361–383, 2017.

[108] A. Athanasiadou, J. Buitelaar, P. Brovedani, O. Chorna, F. Fulceri, A. Guzzetta, and M. L. Scattoni, "Early motor signs of attention-deficit hyperactivity disorder: A systematic review," *European Child & Adolescent Psychiatry*, vol. 29, no. 7, pp. 903–916, 2020.

[109] E. K. Farran, A. Bowler, H. D'Souza, L. Mayall, A. Karmiloff-Smith, E. Sumner, D. Brady, and E. L. Hill, "Is the motor impairment in attention deficit hyperactivity disorder (adhd) a co-occurring deficit or a phenotypic characteristic?," *Advances in Neurodevelopmental Disorders*, vol. 4, no. 3, pp. 253–270, 2020.

[110] D. J. Simmonds, J. J. Pekar, and S. H. Mostofsky, "Meta-analysis of go/no-go tasks demonstrating that fmri activation associated with response inhibition is task-dependent," *Neuropsychologia*, vol. 46, no. 1, pp. 224–232, 2008.

[111] G. M. Lage, L. F. Malloy-Diniz, F. S. Neves, P. H. P. de Moraes, and H. Corrêa, "A kinematic analysis of the association between impulsivity and manual aiming control," *Human movement science*, vol. 31, no. 4, pp. 811–823, 2012.

[112] K. Cahill-Rowley and J. Rose, "Temporal–spatial reach parameters derived from inertial sensors: Comparison to 3d marker-based motion capture," *Journal of biomechanics*, vol. 52, pp. 11–16, 2017.

[113] M. Henry, C. C. Joyal, and P. Nolin, "Development and initial assessment of a new paradigm for assessing cognitive and motor inhibition: the bimodal virtual-reality stroop," *Journal of neuroscience methods*, vol. 210, no. 2, pp. 125–131, 2012.

[114] K. Rubia, T. Russell, S. Overmeyer, M. J. Brammer, E. T. Bullmore, T. Sharma, A. Simmons, S. C. Williams, V. Giampietro, C. M. Andrew, *et al.*, "Mapping motor inhibition: conjunctive brain activations across different versions of go/no-go and stop tasks," *Neuroimage*, vol. 13, no. 2, pp. 250–261, 2001.

[115] J. R. Wessel, "Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm," *Psychophysiology*, vol. 55, no. 3, p. e12871, 2018.

[116] K. M. Trewartha, V. B. Penhune, and K. Z. Li, "Movement kinematics of prepotent response suppression in aging during conflict adaptation," *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 66, no. 2, pp. 185–194, 2011.

[117] R. G. Vaurio, D. J. Simmonds, and S. H. Mostofsky, "Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands," *Neuropsychologia*, vol. 47, no. 12, pp. 2389–2396, 2009.

[118] J. O. Miller and K. Low, "Motor processes in simple, go/no-go, and choice reaction time tasks: a psychophysiological analysis.," *Journal of experimental psychology: Human perception and performance*, vol. 27, no. 2, p. 266, 2001.

[119] A. Dahan, R. Bennet, and M. Reiner, "How long is too long: an individual time-window for motor planning," *Frontiers in human neuroscience*, vol. 13, p. 238, 2019.

[120] L. Wright, J. Lipszyc, A. Dupuis, S. W. Thayapararajah, and R. Schachar, "Response inhibition and psychopathology: A meta-analysis of go/no-go task performance.," *Journal of Abnormal Psychology*, vol. 123, no. 2, p. 429, 2014.

[121] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *Journal of neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[122] L. Iuppariello, G. D'addio, B. Lanzillo, P. Balbi, E. Andreozzi, G. Improta, G. Faiella, and M. Cesarelli, "A novel approach to estimate the upper limb reaching movement in three-dimensional space," *Informatics in Medicine Unlocked*, vol. 15, p. 100155, 2019.

[123] K. J. Wisneski and M. J. Johnson, "Quantifying kinematics of purposeful movements to real, imagined, or absent functional objects: implications for modelling trajectories for robot-assisted adl tasks," *Journal of NeuroEngineering and Rehabilitation*, vol. 4, no. 1, p. 7, 2007.

[124] M. Bourguignon, X. De Tiège, M. O. de Beeck, B. Pirotte, P. Van Bogaert, S. Goldman, R. Hari, and V. Jousmäki, "Functional motor-cortex mapping using corticokinematic coherence," *Neuroimage*, vol. 55, no. 4, pp. 1475–1479, 2011.

[125] M. Bourguignon, V. Jousmäki, M. O. de Beeck, P. Van Bogaert, S. Goldman, and X. De Tiège, "Neuronal network coherent with hand kinematics during fast repetitive hand movements," *Neuroimage*, vol. 59, no. 2, pp. 1684–1691, 2012.

[126] E. Thelen, D. Corbetta, and J. P. Spencer, "Development of reaching during the first year: role of movement speed.," *Journal of experimental psychology: human perception and performance*, vol. 22, no. 5, p. 1059, 1996.

[127] C. Becchio, L. Sartori, M. Bulgheroni, and U. Castiello, "Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement," *Cognition*, vol. 106, no. 2, pp. 894–912, 2008.

[128] A. Dahan and M. Reiner, "Evidence for deficient motor planning in adhd," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[129] M. D. Rapport, S. A. Orban, M. J. Kofler, and L. M. Friedman, "Do programs designed to train working memory, other executive functions, and attention benefit children with adhd? a meta-analytic review of cognitive, aca-

demic, and behavioral outcomes," *Clinical psychology review*, vol. 33, no. 8, pp. 1237–1252, 2013.

[130] S. Cortese, M. Ferrin, D. Brandeis, J. Buitelaar, D. Daley, R. W. Dittmann, M. Holtmann, P. Santosh, J. Stevenson, A. Stringaris, *et al.*, "Cognitive training for attention-deficit/hyperactivity disorder: meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 54, no. 3, pp. 164–174, 2015.

[131] D. Moreau, "Brains and brawn: Complex motor activities to maximize cognitive enhancement," *Educational Psychology Review*, vol. 27, no. 3, pp. 475–482, 2015.

[132] E. Sonuga-Barke and A. Thapar, "The neurodiversity concept: is it helpful for clinicians and scientists?," *The Lancet Psychiatry*, vol. 8, no. 7, pp. 559–561, 2021.

[133] C. Wåhlstedt, L. B. Thorell, and G. Bohlin, "Heterogeneity in adhd: neuropsychological pathways, comorbidity and symptom domains," *Journal of abnormal child psychology*, vol. 37, no. 4, pp. 551–564, 2009.

[134] M. Schweiger and G. M. Marzocchi, "Lo sviluppo delle funzioni esecutive: Uno studio su ragazzi dalla terza elementare alla terza media," *Giornale italiano di psicologia*, vol. 35, no. 2, pp. 353–374, 2008.

[135] D. McIntosh, L. Miller, and V. Shyu, "Development and validation of the short sensory profile (ssp)," *The Sensory Profile: Examiner's Manual; Dunn, W., Ed*, pp. 59–73.

[136] F. Fulceri, A. Narzisi, F. Apicella, G. Balboni, S. Baldini, J. Brocchini, I. Domenici, S. Cerullo, R. Igliozzi, A. Cosenza, *et al.*, "Application of the repetitive behavior scale-revised–italian version–in preschoolers with autism spectrum disorder," *Research in developmental disabilities*, vol. 48, pp. 43–52, 2016.

[137] A. Vallesi, V. N. Lozano, and Á. Correa, "Dissociating temporal preparation processes as a function of the inter-trial interval duration," *Cognition*, vol. 127, no. 1, pp. 22–30, 2013.

[138] P. Ertzgaard, F. Öhberg, B. Gerdle, and H. Grip, "A new way of assessing arm function in activity using kinematic exposure variation analysis and portable inertial sensors–a validity study," *Manual Therapy*, vol. 21, pp. 241–249, 2016.

[139] E. Niechwiej-Szwedo, D. Gonzalez, M. Nouredanesh, and J. Tung, "Evaluation of the leap motion controller during the performance of visually-guided upper limb movements," *PloS one*, vol. 13, no. 3, p. e0193639, 2018.

[140] "Javafx." `https://www.oracle.com/java/technologies/javase/javafx-overview.html`.

[141] A. Angeli, E. L. Piccolomini, and G. Marfia, "Learning about fashion exploiting the big multimedia data," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 48–51, IEEE, 2018.

[142] V. T. Van Hees, L. Gorzelniak, E. C. D. Leon, M. Eder, M. Pias, S. Taherian, U. Ekelund, F. Renström, P. W. Franks, A. Horsch, *et al.*, "Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity," *PloS one*, vol. 8, no. 4, p. e61691, 2013.

[143] E. Domellöf, A. Bäckström, A.-M. Johansson, L. Rönnqvist, C. von Hofsten, and K. Rosander, "Kinematic characteristics of second-order motor planning and performance in 6-and 10-year-old children and adults: Effects of age and task constraints," *Developmental Psychobiology*, vol. 62, no. 2, pp. 250–265, 2020.

[144] E.-J. Wagenmakers and S. Farrell, "Aic model selection using akaike weights," *Psychonomic bulletin & review*, vol. 11, no. 1, pp. 192–196, 2004.

[145] J. C. Douma and J. T. Weedon, "Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression," *Methods in Ecology and Evolution*, vol. 10, no. 9, pp. 1412–1430, 2019.

[146] "R." `https://www.r-project.org/`.

[147] M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *The R Journal*, vol. 9, no. 2, pp. 378–400, 2017.

[148] M. J. Mazerolle, *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)*, 2020. R package version 2.3-1.

[149] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining r2 from generalized linear mixed-effects models," *Methods in ecology and evolution*, vol. 4, no. 2, pp. 133–142, 2013.

[150] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[151] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[152] P. Niemi and R. Näätänen, "Foreperiod and simple reaction time.," *Psychological bulletin*, vol. 89, no. 1, p. 133, 1981.

[153] A. Vallesi and T. Shallice, "Developmental dissociations of preparation over time: deconstructing the variable foreperiod phenomena.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 6, p. 1377, 2007.

[154] G. M. Lage, L. F. Malloy-Diniz, F. S. Neves, L. G. Gallo, A. S. Valentini, and H. Corrêa, "A kinematic analysis of manual aiming control on euthymic bipolar disorder," *Psychiatry research*, vol. 208, no. 2, pp. 140–144, 2013.

[155] S. J. Suskauer, D. J. Simmonds, B. S. Caffo, M. B. Denckla, J. J. Pekar, and S. H. Mostofsky, "fmri of intrasubject variability in adhd: anomalous premotor activity with prefrontal compensation," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 47, no. 10, pp. 1141–1150, 2008.

[156] L. Craighero and G. Rizzolatti, "The premotor theory of attention," in *Neurobiology of attention*, pp. 181–186, Elsevier, 2005.

[157] D. T. Smith and T. Schenk, "The premotor theory of attention: time to move on?," *Neuropsychologia*, vol. 50, no. 6, pp. 1104–1114, 2012.

[158] M. Mari, U. Castiello, D. Marks, C. Marraffa, and M. Prior, "The reach–to–grasp movement in children with autism spectrum disorder," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 358, no. 1430, pp. 393–403, 2003.

[159] S. Glover and P. Dixon, "Semantics affect the planning but not control of grasping," *Experimental Brain Research*, vol. 146, no. 3, pp. 383–387, 2002.

[160] M. M. Demers, N. McNevin, and N. R. Azar, "Adhd and motor control: a review of the motor control deficiencies associated with attention deficit/hyperactivity disorder and current treatment options," *Critical Reviews^{TM} in Physical and Rehabilitation Medicine*, vol. 25, no. 3-4, 2013.

[161] U. Castiello, "The neuroscience of grasping," *Nature Reviews Neuroscience*, vol. 6, no. 9, pp. 726–736, 2005.

[162] O. E. Krigolson, D. Cheng, and G. Binsted, "The role of visual processing in motor learning and control: Insights from electroencephalography," *Vision Research*, vol. 110, pp. 277–285, 2015.

[163] V. Boulenger, B. Y. Silber, A. C. Roy, Y. Paulignan, M. Jeannerod, and T. A. Nazir, "Subliminal display of action words interferes with motor planning: a combined eeg and kinematic study," *Journal of Physiology-Paris*, vol. 102, no. 1-3, pp. 130–136, 2008.

[164] H. M. Eckert and D. H. Eichorn, "Developmental variability in reaction time," *Child Development*, pp. 452–458, 1977.

[165] G. Mirabella, "Inhibitory control and impulsive responses in neurodevelopmental disorders," *Developmental Medicine & Child Neurology*, 2020.

[166] D. S. Aichert, N. M. Wöstmann, A. Costa, C. Macare, J. R. Wenig, H.-J. Möller, K. Rubia, and U. Ettinger, "Associations between trait impulsivity and prepotent response inhibition," *Journal of clinical and experimental neuropsychology*, vol. 34, no. 10, pp. 1016–1032, 2012.

[167] F. Uzefovsky, C. Allison, P. Smith, and S. Baron-Cohen, "Brief report: the go/no-go task online: inhibitory control deficits in autism in a large sample," *Journal of Autism and Developmental Disorders*, vol. 46, no. 8, pp. 2774–2779, 2016.

[168] M. Mosconi, M. Kay, A.-M. D'cruz, A. Seidenfeld, S. Guter, L. Stanford, and J. Sweeney, "Impaired inhibitory control is associated with higher-order repetitive behaviors in autism spectrum disorders," *Psychological medicine*, vol. 39, no. 9, p. 1559, 2009.

[169] A. Miyake and N. P. Friedman, "The nature and organization of individual differences in executive functions: Four general conclusions," *Current directions in psychological science*, vol. 21, no. 1, pp. 8–14, 2012.

[170] A. Bisio, L. Pedullà, L. Bonzano, A. Tacchino, G. Brichetto, and M. Bove, "The kinematics of handwriting movements as expression of cognitive and sensorimotor impairments in people with multiple sclerosis," *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.

[171] S. Forti, A. Valli, P. Perego, M. Nobile, A. Crippa, and M. Molteni, "Motor planning and control in autism. a kinematic analysis of preschool children," *Research in Autism Spectrum Disorders*, vol. 5, no. 2, pp. 834–842, 2011.

[172] A. Schröter, R. Mergl, K. Bürger, H. Hampel, H.-J. Möller, and U. Hegerl, "Kinematic analysis of handwriting movements in patients with alzheimer's

disease, mild cognitive impairment, depression and healthy subjects," *Dementia and geriatric cognitive disorders*, vol. 15, no. 3, pp. 132–142, 2003.

[173] A. Smiley-Oyen, K. Lowry, and J. Kerr, "Planning and control of sequential rapid aiming in adults with parkinson's disease," *Journal of Motor Behavior*, vol. 39, no. 2, pp. 103–114, 2007.

[174] E. M. Mahone, D. Bridges, C. Prahme, and H. S. Singer, "Repetitive arm and hand movements (complex motor stereotypies) in children," *The Journal of pediatrics*, vol. 145, no. 3, pp. 391–395, 2004.

[175] E. M. Mahone, M. Ryan, L. Ferenc, C. Morris-Berry, and H. S. Singer, "Neuropsychological function in children with primary complex motor stereotypies," *Developmental Medicine & Child Neurology*, vol. 56, no. 10, pp. 1001–1008, 2014.

[176] K. M. Harris, E. M. Mahone, and H. S. Singer, "Nonautistic motor stereotypies: clinical features and longitudinal follow-up," *Pediatric neurology*, vol. 38, no. 4, pp. 267–272, 2008.

[177] D. Ghosh, P. V. Rajan, and G. Erenberg, "A comparative study of primary and secondary stereotypies," *Journal of child neurology*, vol. 28, no. 12, pp. 1562–1568, 2013.

[178] E. Houdayer, J. Walthall, B. A. Belluscio, S. Vorbach, H. S. Singer, and M. Hallett, "Absent movement-related cortical potentials in children with primary motor stereotypies," *Movement Disorders*, vol. 29, no. 9, pp. 1134–1140, 2014.

[179] G. Mirabella, C. Mancini, F. Valente, and F. Cardona, "Children with primary complex motor stereotypies show impaired reactive but not proactive inhibition," *Cortex*, vol. 124, pp. 250–259, 2020.

[180] N. J. Brierley, C. G. McDonnell, K. Parks, S. E. Schulz, T. C. Dalal, E. Kelley, E. Anagnostou, R. Nicolson, S. Georgiades, J. Crosbie, *et al.*, "Factor structure of repetitive behaviors across autism spectrum disorder and

attention-deficit/hyperactivity disorder," *Journal of Autism and Developmental Disorders*, vol. 51, no. 10, pp. 3391–3400, 2021.

[181] J. Jang, J. L. Matson, L. W. Williams, K. Tureck, R. L. Goldin, and P. E. Cervantes, "Rates of comorbid symptoms in children with asd, adhd, and comorbid asd and adhd," *Research in developmental disabilities*, vol. 34, no. 8, pp. 2369–2378, 2013.

[182] E. Sokolova, A. M. Oerlemans, N. N. Rommelse, P. Groot, C. A. Hartman, J. C. Glennon, T. Claassen, T. Heskes, and J. K. Buitelaar, "A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (adhd) and autism spectrum disorder (asd)," *Journal of autism and developmental disorders*, vol. 47, no. 6, pp. 1595–1604, 2017.

[183] S. Faja and L. Nelson Darling, "Variation in restricted and repetitive behaviors and interests relates to inhibitory control and shifting in children with autism spectrum disorder," *Autism*, vol. 23, no. 5, pp. 1262–1272, 2019.

[184] A. Fetta, E. Carati, L. Moneti, V. Pignataro, M. Angotti, M. C. Bardasi, D. M. Cordelli, E. Franzoni, and A. Parmeggiani, "Relationship between sensory alterations and repetitive behaviours in children with autism spectrum disorders: A parents' questionnaire based study," *Brain sciences*, vol. 11, no. 4, p. 484, 2021.

[185] V. T. Shimizu, O. F. Bueno, and M. C. Miranda, "Sensory processing abilities of children with adhd," *Brazilian journal of physical therapy*, vol. 18, pp. 343–352, 2014.

[186] L. M. Little, E. Dean, S. Tomchek, and W. Dunn, "Sensory processing patterns in autism, attention deficit hyperactivity disorder, and typical development," *Physical & occupational therapy in pediatrics*, vol. 38, no. 3, pp. 243–254, 2018.

[187] A. Fuermaier, P. Hüpen, S. M. De Vries, M. Müller, F. M. Kok, J. Koerts, J. Heutink, L. Tucha, M. Gerlach, and O. Tucha, "Perception in attention

deficit hyperactivity disorder," *ADHD Attention Deficit and Hyperactivity Disorders*, vol. 10, no. 1, pp. 21–47, 2018.

[188] I. Valori, L. Carnevali, and T. Farroni, "Agency and reward across development and in autism: A free-choice paradigm," *Plos one*, vol. 18, no. 4, p. e0284407, 2023.

[189] M. R. Peres, *The concise Focal encyclopedia of photography: from the first photo on paper to the digital revolution.* CRC Press, 2014.

[190] E. Serafinelli, *Digital life on Instagram: New social communication of photography.* Emerald Group Publishing, 2018.

[191] W. Yin, T. Mei, C. W. Chen, and S. Li, "Socialized mobile photography: Learning to photograph with social context via mobile devices," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 184–200, 2013.

[192] J. Chen and H. Wang, "Guest editorial: Big data infrastructure i," *IEEE Transactions on Big Data*, vol. 4, no. 2, pp. 148–149, 2018.

[193] E. Borcoci, D. Negru, and C. Timmerer, "A novel architecture for multimedia distribution based on content-aware networking," in *2010 Third International Conference on Communication Theory, Reliability, and Quality of Service*, pp. 162–168, IEEE, 2010.

[194] B. Rainer, S. Petscharnig, C. Timmerer, and H. Hellwagner, "Statistically indifferent quality variation: An approach for reducing multimedia distribution cost for adaptive video streaming services," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 849–860, 2016.

[195] W. Zhang, T. Yao, S. Zhu, and A. E. Saddik, "Deep learning-based multimedia analytics: A review," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, Jan. 2019.

[196] J. Lemley, S. Bazrafkan, and P. Corcoran, "Deep learning for consumer devices and services: Pushing the limits for machine learning, artificial intelli-

gence, and computer vision.," *IEEE Consumer Electronics Magazine*, vol. 6, no. 2, pp. 48–56, 2017.

[197] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Deep metric learning with density adaptivity," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1285–1297, 2019.

[198] X. Chen, D. Liu, Z. Xiong, and Z.-J. Zha, "Learning and fusing multiple user interest representations for micro-video and movie recommendations," *IEEE Transactions on Multimedia*, 2020.

[199] F. Vaccaro, M. Bertini, T. Uricchio, and A. D. Bimbo, "Image retrieval using multi-scale cnn features pooling," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pp. 311–315, 2020.

[200] M. Roccetti, L. Casini, G. Delnevo, V. Orrù, N. Marchetti, and Nicolò, "Potential and limitations of designing a deep learning model for discovering new archaeological sites: A case with the mesopotamian floodplain," in *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 216–221, 2020.

[201] G. Mitman and K. Wilder, *Documenting the world: film, photography, and the scientific record*. Univ. of Chicago Press, 2016.

[202] MoMA, "Vernacular photography." `https://www.moma.org/collection/terms/vernacular-photography`, 2020.

[203] M. Sandbye, "Looking at the family photo album: a resumed theoretical discussion of why and how," *Journal of Aesthetics & Culture*, vol. 6, no. 1, p. 25419, 2014.

[204] D. Calanca, "Italians posing between public and private. theories and practices of social heritage," *Almatourism-Journal of Tourism, Culture and Territorial Development*, vol. 2, no. 3, pp. 1–9, 2011.

[205] N. Black, "Why we need qualitative research.," *Journal of epidemiology and community health*, vol. 48, no. 5, p. 425, 1994.

[206] J. Michell, *Measurement in psychology: A critical history of a methodological concept*, vol. 53. Cambridge University Press, 1999.

[207] A. Tashakkori and J. W. Creswell, "The new era of mixed methods," 2007.

[208] E. P. Baumer, D. Mimno, S. Guha, E. Quan, and G. K. Gay, "Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?," *Journal of the Association for Information Science and Technology*, vol. 68, no. 6, pp. 1397–1410, 2017.

[209] J. Radford and K. Joseph, "Theory in, theory out: the uses of social theory in machine learning for social science," *Frontiers in big Data*, vol. 3, p. 18, 2020.

[210] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang, "Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 325–336, 2020.

[211] M. K. Scheuerman, A. Hanna, and E. Denton, "Do datasets have politics? disciplinary values in computer vision dataset development," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–37, 2021.

[212] L. T. Choy, "The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches," *IOSR Journal of Humanities and Social Science*, vol. 19, no. 4, pp. 99–104, 2014.

[213] J. Prosser, "The status of image-based research," *Image-based research: A sourcebook for qualitative researchers*, pp. 97–112, 1998.

[214] D. Calanca, "Album di famiglia. autorappresentazioni tra pubblico e privato (1870-1950).," *Storia e Futuro - NÂ° 8-9*, 2005.

[215] A. Molina, P. Riba, L. Gomez, O. Ramos-Terrades, and J. Lladós, "Date estimation in the wild of scanned historical photos: An image retrieval approach," *arXiv preprint arXiv:2106.05618*, 2021.

[216] S. Ginosar, K. Rakelly, S. Sachs, B. Yin, and A. A. Efros, "A century of portraits: A visual historical record of american high school yearbooks," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–7, 2015.

[217] T. Salem, S. Workman, Z. M, and N. Jacobs, "Analyzing human appearance as a cue for dating images," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, IEEE, 2016.

[218] E. Müller, M. Springstein, and R. Ewerth, ""When was this picture taken?"–image date estimation in the wild," in *European Conference on Information Retrieval*, pp. 619–625, Springer, 2017.

[219] E. Coburn, E. Lanzi, E. O'Keefe, R. Stein, and A. Whiteside, "The cataloging cultural objects experience: Codifying practice for the cultural heritage community," *IFLA journal*, vol. 36, no. 1, pp. 16–29, 2010.

[220] S. Barba, F. Fiorillo, P. Ortiz Coder, S. D'auria, E. De Feo, *et al.*, "An application for cultural heritage in erasmus placement. surveys and 3d cataloguing archaeological finds in merida (spain)," 2011.

[221] D. Rosner, M. Roccetti, and G. Marfia, "The digitization of cultural practices," *Commun. ACM*, vol. 57, pp. 82–87, June 2014.

[222] M. Lincoln, J. Corrin, E. Davis, and S. B. Weingart, "Campi: Computer-aided metadata generation for photo archives initiative," 2020.

[223] L. T. T Arnold, "Distant viewing: analyzing large visual corpora," *Digital Scholarship in the Humanities*, vol. 34, no. Supplement_1, pp. i3–i16, 2019.

[224] M. Wevers and T. Smits, "The visual digital turn: Using neural networks to study historical images," *Digital Scholarship in the Humanities*, vol. 35, no. 1, pp. 194–207, 2020.

[225] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, "Deep residual learning for image recognition," 2015.

[226] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015.

[227] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.

[228] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[229] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, pp. 10347–10357, PMLR, 2021.

[230] J. Saldaña, *The coding manual for qualitative researchers*. sage, 2021.

[231] Amazon, "Amazon sagemaker ground truth." `https://aws.amazon.com/it/sagemaker/groundtruth/`, 2021.

[232] Google, "Ai platform data labeling service." `https://cloud.google.com/ai-platform/data-labeling/docs`, 2021.

[233] F. Palermo, J. Hays, and A. A. Efros, "Dating historical color images," in *European Conference on Computer Vision*, pp. 499–512, Springer, 2012.

[234] F. Basura, M. Damien, R. Khan, and T. Tuytelaars, "Color features for dating historical color images," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 2589–2593, IEEE, 2014.

[235] P. Sorcinelli, "Imago. laboratorio di ricerca storica e di documentazione iconografica sulla condizione giovanile nel xx secolo," *Rivista di storia e storiografia*, vol. 5, pp. 200–202, Nov. 2004.

[236] D. Calanca, "Percorsi di storia della famiglia," *Rivista di storia e storiografia*, vol. 5, pp. 203–210, Nov. 2004.

[237] D. Calanca, "Fotografie amatoriali e fotografie professionali nell'italia del boom economico," *Storia e Futuro - NÂ° 12*, 2006.

[238] J. Scott and G. Marshall, *A dictionary of sociology*. Oxford University Press, USA, 2009.

[239] M. A. Cabrera, "On language, culture, and social action," *History and Theory*, vol. 40, no. 4, pp. 82–100, 2001.

[240] M. Á. Cabrera, *Postsocial history: An introduction*. Lexington Books, 2004.

[241] L. Criscenti, G. D'autilia, and G. D. Luna, *L'Italia del Novecento: Le fotografie e la storia*. Giulio Einaudi editore, 2005.

[242] P. Bourdieu, "On the family as a realized category," *Theory, culture & society*, vol. 13, no. 3, pp. 19–26, 1996.

[243] K. J. Enns and M. J. Martin, "Gendering agricultural education: A study of historical pictures of women in the agricultural education magazine.," *Journal of Agricultural Education*, vol. 56, no. 3, pp. 69–89, 2015.

[244] L. Bosi and H. Reiter, "Historical methodologies," *Methodological practices in social movement research*, pp. 117–43, 2014.

[245] E. S. Clemens and M. D. Hughes, "Recovering past protest: Historical research on social movements," *Methods of social movement research*, vol. 16, pp. 201–230, 2002.

[246] K. Bentein, "Minor complementation patterns in post-classical greek (i–vi ad): A socio-historical analysis of a corpus of documentary papyri," *Symbolae Osloenses*, vol. 89, no. 1, pp. 104–147, 2015.

[247] C. Schreiber, "The construction of 'female citizens': a socio-historical analysis of girls' education in luxembourg," *Educational Research*, vol. 56, no. 2, pp. 137–154, 2014.

[248] R. Franzosi, "Narrative as data: linguistic and statistical tools for the quantitative study of historical events," *International review of social history*, vol. 43, no. S6, pp. 81–104, 1998.

[249] T. J. Sejnowski, *The deep learning revolution*. MIT press, 2018.

[250] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and timeseries forecasting," in *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*, pp. 1–6, 2014.

[251] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[252] T Nguyen, "Yolo face implementation." `https://github.com/sthanhng/yoloface`, 2018. Online; accessed 3 August 2020.

[253] J Redmon, "YOLO: Real Time Object Detection." `https://github.com/pjreddie/darknet/wiki/YOLO:-Real-Time-Object-Detection`, 2019. Online; accessed 3 August 2020.

[254] K. Zhang, W. Zuo, and L. Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising." IEEE Transactions on Image Processing, 2018.

[255] J. T. S Paris, P Kornprobst and F. Durand, "A gentle introduction to bilateral filtering and its applications," in *ACM SIGGRAPH 2007 Courses*, SIGGRAPH '07, (New York, NY, USA), pp. 1–es, Association for Computing Machinery, 2007.

[256] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," 2018.

[257] K. Zhang, "Image restoration toolbox." `https://github.com/cszn/KAIR`, 2019.

[258] J. Deng, W. Dong, R. Socher, L. J. Li, K Li, and L Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[259] E. Culurciello, "Neural network architectures." `https://towardsdatascience.com/neural-network-architectures-156e5bad51ba`, 2021.

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

[261] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A survey on visual transformer," *arXiv preprint arXiv:2012.12556*, 2020.

[262] T. H. Phan and K. Yamamoto, "Resolving class imbalance in object detection with weighted cross entropy losses," 2020.

[263] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, Oct 2019.

[264] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[265] S. Gundle and M. Guani, "L'americanizzazione del quotidiano. televisione e consumismo nell'italia degli anni cinquanta," *Quaderni storici*, pp. 561–594, 1986.

[266] W. post, "How america became italian." `https://www.washingtonpost.com/opinions/how-america-became-italian/2015/10/09/4c93b1be-6ddd-11e5-9bfe-e59f5e244f92_story.html?utm_term=.5a515dec12c5&noredirect=on`, 2022.

[267] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[268] D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, "Mobile augmented reality survey: From where we are to where we go," *Ieee Access*, vol. 5, pp. 6917–6950, 2017.

[269] L. Stacchio, S. Hajahmadi, and G. Marfia, "Preserving family album photos with the hololens 2," in *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 643–644, IEEE, 2021.

[270] G. Strezoski and M. Worring, "Omniart: Multi-task deep learning for artistic data analysis," 2017.

[271] B. C. G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, and D. S. Weld, "The newspaper navigator dataset: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america," *arXiv preprint arXiv:2005.01583*, 2020.

[272] Z. Lu, M. Annett, M. Fan, and D. Wigdor, "" i feel it is my responsibility to stream" streaming and engaging with intangible cultural heritage through livestreaming," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[273] J. Jetter, J. Eimecke, and A. Rese, "Augmented reality tools for industrial applications: What are potential key performance indicators and who benefits?," *Computers in Human Behavior*, vol. 87, pp. 18–33, 2018.

[274] G. Marfia, M. Roccetti, G. Matteucci, and A. Marcomini, "Technoculture of handcraft: Fine gesture recognition for haute couture skills preservation and transfer in italy," in *ACM SIGGRAPH 2012 Posters*, SIGGRAPH '12, (New York, NY, USA), Association for Computing Machinery, 2012.

[275] G. Marfia, M. Roccetti, A. Marcomini, C. Bertuccioli, and G. Matteucci, "Reframing haute couture handcraftship: how to preserve artisans' abilities with gesture recognition," in *International Conference on Advances in Computer Entertainment Technology*, pp. 437–444, Springer, 2012.

[276] L. Penco, F. Serravalle, G. Profumo, and M. Viassone, "Mobile augmented reality as an internationalization tool in the 'made in italy' food and beverage industry," *Journal of Management and Governance*, vol. 25, no. 4, pp. 1179–1209, 2021.

[277] A. Sonderegger, D. Ribes, N. Henchoz, and E. Groves, "Food talks: visual and interaction principles for representing environmental and nutritional food information in augmented reality," in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 98–103, IEEE, 2019.

[278] V. Gundimeda, R. S. Murali, R. Joseph, and N. N. Babu, "An automated computer vision system for extraction of retail food product metadata," in *First International Conference on Artificial Intelligence and Cognitive Computing*, pp. 199–216, Springer, 2019.

[279] B. Hu, N. Zhou, Q. Zhou, X. Wang, and W. Liu, "Diffnet: a learning to compare deep network for product recognition," *IEEE Access*, vol. 8, pp. 19336–19344, 2020.

[280] M. Lin, L. Ma, and B. Yu, "An efficient and light-weight detector for wine bottle defects," in *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 957–962, IEEE, 2020.

[281] TinEye, "Wineengine is image recognition for the beverage industry." `https://services.tineye.com/WineEngine`, 2021.

[282] livingwinelabels, "livingwinelabels." `https://www.livingwinelabels.com/`, 2021.

[283] Vivino, "Vivino." `https://www.vivino.com/`, 2021.

[284] PTC, "Vivino and Vuforia's Image Recognition Solution Make a Great Pairing." `https://www.ptc.com/en/case-studies/vivino`, 2022.

[285] T. Gebru, O. Hazi, and V. Yeh, "Mobile wine label recognition,"

[286] M.-Y. Wu, J.-H. Lee, and S.-W. Kuo, "A hierarchical feature search method for wine label image recognition," in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 568–572, IEEE, 2015.

[287] J.-M. Jung, H.-J. Yang, S.-H. Kim, G.-S. Lee, and S.-H. Kim, "Wine label recognition system using image similarity," *The Journal of the Korea Contents Association*, vol. 11, no. 5, pp. 125–137, 2011.

[288] X. Li, J. Yang, and J. Ma, "Cnn-sift consecutive searching and matching for wine label retrieval," in *International Conference on Intelligent Computing*, pp. 250–261, Springer, 2019.

[289] J. O. Álvarez Márquez and J. Ziegler, "Improving the shopping experience with an augmented reality-enhanced shelf," *Mensch und Computer 2017-Workshopband*, 2017.

[290] Vuforia, "Vuforia SDK." `https://developer.vuforia.com/downloads/SDK`, 2022.

[291] I. S. Na, Y. J. Chen, and S. H. Kim, "Automatic segmentation of product bottle label based on grabcut algorithm," *International Journal of Contents*, vol. 10, no. 4, pp. 1–10, 2014.

[292] S. Čakić, T. Popović, S. Šandi, S. Krčo, and A. Gazivoda, "The use of tesseract ocr number recognition for food tracking and tracing," in *2020 24th International Conference on Information Technology (IT)*, pp. 1–4, IEEE, 2020.

[293] D. Ungureanu, F. Bogo, S. Galliani, P. Sama, X. Duan, C. Meekhof, J. Stühmer, T. J. Cashman, B. Tekin, J. L. Schönberger, *et al.*, "Hololens 2 research mode as a tool for computer vision research," *arXiv preprint arXiv:2008.11239*, 2020.

[294] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015.

[295] "Advances in computer vision," *Advances in Intelligent Systems and Computing*, 2020.

[296] Ultralytics, "Yolo v5." `https://github.com/ultralytics/yolov5`, 2021. Online; accessed 06 June 2021.

[297] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2015.

[298] P. Salomoni, C. Prandi, M. Roccetti, L. Casanova, L. Marchetti, and G. Marfia, "Diegetic user interfaces for virtual environments with hmds: a user experience study with oculus rift," *Journal on Multimodal User Interfaces*, vol. 11, 01 2017.

[299] Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," *Commun. ACM*, vol. 35, pp. 379–383, 08 2003.

[300] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[301] A. Rese, D. Baier, A. Geyer-Schulz, and S. Schreiber, "How augmented reality apps are accepted by consumers: A comparative analysis using scales and opinions," *Technological Forecasting and Social Change*, vol. 124, pp. 306–319, 2017.

[302] J. Pallant, "Spss survival manual 4th edition," *Everbest Printing*, 2011.

[303] C. Phelan and J. Wren, "Exploring reliability in academic assessment," *UNI Office of Academic Assessment*, pp. 92005–2006, 2006.

[304] F. Tao, H. Zhang, A. Liu, and A. Y. Nee, "Digital twin in industry: State-of-the-art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2018.

[305] A. Rasheed, O. San, and T. Kvamsdal, "Digital twin: Values, challenges and enablers from a modeling perspective," *Ieee Access*, vol. 8, pp. 21980–22012, 2020.

[306] S. M. Sepasgozar, "Digital twin and web-based virtual gaming technologies for online education: A case of construction management and engineering," *Applied Sciences*, vol. 10, no. 13, p. 4678, 2020.

[307] H. Elayan, M. Aloqaily, and M. Guizani, "Digital twin for intelligent context-aware iot healthcare systems," *IEEE Internet of Things Journal*, 2021.

[308] A. Mukhopadhyay, G. R. Reddy, K. S. SALUJA, S. Ghosh, A. Peña-Rios, G. GOPAL, and P. Biswas, "Virtual-reality-based digital twin of office spaces with social distance measurement feature," *Virtual Reality & Intelligent Hardware, XXXX, XX (XX)*, pp. 1–21, 2021.

[309] Y. Chen, G. M. Lee, L. Shu, and N. Crespi, "Industrial internet of things-based collaborative sensing intelligence: framework and research challenges," *Sensors*, vol. 16, no. 2, p. 215, 2016.

[310] W. Li, W.-j. Wu, H.-m. Wang, X.-q. Cheng, H.-j. Chen, Z.-h. Zhou, and R. Ding, "Crowd intelligence in ai 2.0 era," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 15–43, 2017.

[311] X. Niu and S. Qin, "Integrating crowd-/service-sourcing into digital twin for advanced manufacturing service innovation," *Advanced Engineering Informatics*, vol. 50, p. 101422, 2021.

[312] L. Bononi, L. Donatiello, D. Longo, M. Massari, F. Montori, L. Stacchio, and G. Marfia, "Digital twin collaborative platforms: Applications to humans-in-the-loop crafting of urban areas," *IEEE Consumer Electronics Magazine*, 2022.

[313] A. Ismail, "Benchmark mobile device cameras," 2021.

[314] G. Chandan, A. Jain, H. Jain, and Mohana, "Real time object detection and tracking using deep learning and opencv," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1305–1308, 2018.

[315] EasyOcr, "JadedAI." `https://github.com/JaidedAI/EasyOCR`, 2020.

[316] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," 2016.

[317] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 986–996, Springer, 2003.

[318] M. R. Abbasifard, B. Ghahremani, and H. Naderi, "A survey on nearest neighbor search methods," *International Journal of Computer Applications*, vol. 95, no. 25, 2014.

[319] M. Krichenbauer, G. Yamamoto, T. Taketom, C. Sandor, and H. Kato, "Augmented reality versus virtual reality for 3d object manipulation," *IEEE*

Transactions on Visualization and Computer Graphics, vol. 24, no. 2, pp. 1038–1048, 2018.

[320] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54 – 71, 2019.

[321] S. Charters, L. Lockshin, and T. Unwin, "Consumer responses to wine bottle back labels," *Journal of Wine Research*, vol. 10, no. 3, pp. 183–195, 1999.

[322] J. M. Alston and D. Gaeta, "Reflections on the political economy of european wine appellations," *Italian Economic Journal*, vol. 7, no. 2, pp. 219–258, 2021.

[323] Camera di Commercio Molise, "Guida etichettature vino." `https://www.molise.camcom.gov.it/sites/default/files/guida_etichettatura_vino.pdf`, 2016.

[324] Michele A. Fino, "Questione di Etichetta." `https://www.spazioprever.it/salabar/vino/pdf/Questione_di_etichetta.pdf`, 2013.

[325] Vittorio Portinari, "Elementi di Legislazione Vitivinicola: le norme per l'etichettatura e la tracciabilità dei vini." `http://www.sardegnaagricoltura.it/documenti/14_43_20160531144229.pdf`, 2016.

[326] Federdoc, "I vini italiani a denominazione d'origine 2020." `https://www.federdoc.com/new/wp-content/uploads/2020/06/vini_italiani_denominazione_origine_2020.pdf`, 2021.

[327] A. Singh, K. Bacchuwar, and A. Bhasin, "A survey of ocr applications," *International Journal of Machine Learning and Computing*, vol. 2, no. 3, p. 314, 2012.

[328] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9365–9374, 2019.

[329] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[330] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4715–4723, 2019.

[331] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, 2006.

[332] K. Smelyakov, A. Chupryna, D. Darahan, and S. Midina, "Effectiveness of modern text recognition solutions and tools for common data sources," in *CEUR Workshop Proceedings*, pp. 154–165, 2021.

[333] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, Soviet Union, 1966.

[334] R. Bansal, G. Raj, and T. Choudhury, "Blur image detection using laplacian operator and open-cv," in *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, pp. 63–67, 2016.

[335] F. Zhan and S. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2059–2068, 2019.

[336] A. Lat and C. Jawahar, "Enhancing ocr accuracy with super resolution," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3162–3167, IEEE, 2018.

[337] P. Follmann, B. Drost, and T. Böttger, "Acquire, augment, segment and enjoy: Weakly supervised instance segmentation of supermarket products," in *Pattern Recognition* (T. Brox, A. Bruhn, and M. Fritz, eds.), (Cham), pp. 363–376, Springer International Publishing, 2019.