UNIVERSITÀ degli STUDI di CATANIA

DIPARTIMENTO DI INGEGNERIA ELETTRICA, ELETTRONICA E INFORMATICA

DOTTORATO DI RICERCA IN INGEGNERIA INFORMATICA E DELLE TELECOMUNICAZIONI

XXVIII CICLO

Ph.D. Thesis

MULTIFACETED ANALYSIS FOR MEDICAL DATA UNDERSTANDING: FROM DATA ACQUISITION TO MULTIDIMENSIONAL SIGNAL PROCESSING TO KNOWLEDGE DISCOVERY

ING. ISAAK KAVASIDIS

Coordinatore	Tutor
Chiar.ma Prof.ssa	Chiar.ma Prof.ssa
D. CARCHIOLO	D. GIORDANO

to my family

ABSTRACT

Large quantities of medical data are routinely generated each day in the form of text, images and time signals, making evident the need to develop new methodologies not only for the automatization of the processing and management of such data, but also for the deeper understanding of the concepts hidden therein. The main problem that arises is that the acquired data cannot always be in an appropriate state or quality for quantitative analysis, and further processing is often necessary in order to enable automatic processing and management as well as to increase the accuracy of the results. Also, given the multimodal nature of medical data uniform approaches no longer apply and specific algorithm pipelines should be conceived and developed for each case.

In this dissertation we tackle some of the problems that occur in the medical domain regarding different data modalities and an attempt to understand the meaning of these data is made. These problems range from cortical brain signal acquisition and processing to X-Ray image analysis to text and genomics data-mining and subsequent knowledge discovery.

CONTENTS

CHAPTER

ONE

INTRODUCTION

The medical domain is characterized by a profound multimodality in the processing of data. Text, images and time signals are the predominant data formats that are used during screening and diagnostic processes. The main problem that arises though, is that the acquired data cannot always be in an appropriate state or quality for quantitative analysis and further processing is often necessary in order not only to enable automatic processing and management, but also to increase the accuracy of the results.

For example, Electroencephalography, is one of the most common examinations that is used for monitoring brain activity and for diagnosing conditions of the central and peripheral nervous system. Before an EEG signal is considered, it goes under heavy filtering for removing unwanted artifacts and noise that would render the signal unusable. Another example is medical imaging data. Magnetic resonance imaging, computer tomography, X-Ray and echography images all need an

initial preprocessing step to remove undesired information and to enhance the quality of the image by using image processing techniques.

Quality of the medical information is not the only desired outcome. Manageability of the resources is another requested ability of medical institutions worldwide. For instance, thousands of medical reports are generated daily by medical institutions, but very often, are hand written and when digital systems are used for their management, they do not always use common formats and sharing or migrating text documents between medical institutions require a further integration step that most often comes at prohibitive costs.

Numerous research groups produce resources, but what happens when not identical, but relevant, research tracks are combined in an automatic and exploratory way? Is it possible to bring together resources and discover hidden knowledge by analyzing their common parts? Even more interestingly, given the widespread diffusion of medical information, is it possible that disparate pieces of information that were not considered before, if combined, could provoke a breakthrough in biomedical research?

During my Ph.D. course, I focused my research efforts on dealing with the aforementioned problems, and for this reason I identified 4 diverse problems of the medical domain that should be dealt with by different approaches. In the following subsection a brief description of such problems is given.

1.1 Structure of this Dissertation

This dissertation is organized in seven chapters (including this introduction) as follows:

Chapter 2 presents a method for generating automatically summaries based on patients' reports by employing Natural Language Processing techniques. This method aims at creating structured text, based on unstructured text, that can be easily managed, stored and shared.

Chapter 3 presents a method and a suite of tools for assisting neurophysiologists to create and execute large-scale experiments based on paired-pulse Transcranial Magnetic Stimulation. The tools contained in the suite cover completely the life-cycle of large-scale medical experiments (i.e. experiment definition, stimuli administration, signal acquisition and statistical analysis).

Chapter 4 presents a method and a tool for assessing the skeletal bone age of an individual based on X-Ray images of the left hand. Such a method, is important in many contexts ranging from legal rights assessment, to devolepmental disorders. Machine learning techniques, and in particular Hidden Markov Models, deal with the classification task, achieving remarkable results.

Chapter 5 describes a tool *(BioCloud)* that conducts knowledge discovery in the biomedical domain by processing a multitude of sources and data formats. It is capable of processing large quantities of scientific literature papers, online genomics databases and disease related databases to establish and verify relations between genes, proteins and biological processes that lead to disease. Given the high volumes of data that the tool needs to process, the whole processing

and data flow has been parallelized and deployed as a cloud service in order to exploit the high throughput these paradigms can offer.

Finally, in Chapter 6 conclusions are drawn and future directions are given.

Each chapter is independent with each other, as they deal with completely different problems of the medical domain.

CHAPTER

TWO

TEXT PROCESSING: CREATING SUMMARIES OF UNSTRUCTURED MEDICAL RECORDS

In this chapter we present a system for automatic generation of summaries of patients' unstructured medical reports. The system employs Natural Language Processing techniques in order to determine the most interesting points and uses the MetaMap module for recognizing the medical concepts in a medical report. Afterwards the sentences that do not contain interesting concepts are removed and a summary is generated which contains URL links to the Linked Life Data pages of the identified medical concepts, enabling both medical doctors and patients to further explore what is reported in. Such integration also allows the tool to interface with other semantic web-based applications. The performance of the tool were also evaluated, achieving remarkable results in sentence identification, polarity detection and concept recognition. Moreover, the accuracy of the generated summaries was

evaluated by five medical doctors, proving that the summaries keep the same relevant information as the medical reports, despite being much more concise.

2.1 Introduction

Every day a large amount of medical reports, in the form of free text (i.e. not structured according to a logical scheme) is generated. Not possessing any structural information hampers the ability of automatic document digitization and analysis and subsequently all the applications that could be built upon these. The information included in the text can be deductible only through reading. The adoption of free text documents is done mainly due to the doctors' lack of time, who have to write reports quickly, or due to hospitals' internal procedures or traditions. Moreover, the readability of these documents could become a problem as it may not be easy for the reader to pinpoint the most important parts.

The medical domain suffers particularly by an overload of information and rapid access to key information is of crucial importance to health professionals for decision making. For instance, a concise and synthetic representation of medical reports (i.e. a summary), could serve to create a precise list of what was performed by the health organization and derive an automatic method for calculating hospitalization costs. Given the plethora in number and diversity of sources of medical documents, the purpose of summarization is to make users able to assimilate and easily determine the contents of a document, and then quickly determine the key points of it. In particular, as reported in

[1]: *"A summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that"*, but also denotes its most important challenge: *"Identifying the information segments at the expense of the rest is the main challenge in summarization"*. Generating summaries, however, is not trivial as it implies a deep understanding of the underlying semantics. This is even more challenging in the medical domain since medical reports include a highly specialized vocabulary, words in upper and lowercase letters and numbers that require ad-hoc tokenization. These problems urged the development of domain-specific resources such as PubMed/MEDLINE and PubMedCentral[1], ontologies and other semantic lexical resources, such as Gene Ontology[2] and Unified Medical Language System *(UMLS)*[3], and annotated databases, such as Entrez Gene[4] which are used heavily by a variety of text mining applications.

The objectives of the work presented herein is 1) to create automatically a summary that conveys the key points of medical reports and 2) to provide a tool for annotating the medical concepts found in the text with Linked Life Data (LLD)[5], so that the doctors or the patients can explore further what is being reported and also enable interoperability with other semantic web-enabled applications.

The remainder of the chapter is as follows: the next section briefly

[1]http://www.ncbi.nlm.nih.gov/pubmed
[2]http://www.geneontology.org/
[3]http://www.nlm.nih.gov/research/umls/
[4]http://www.ncbi.nlm.nih.gov/gene/
[5]http://linkedlifedata.com/

presents related works, while Section 2.3 describes the method in detail and in Section 2.4 a performance evaluation of the system is carried out.

2.2 Related Work

Text summarization of medical documents was brought to the attention of the scientific community due to the tremendous growth of information that are available to physicians and researchers: the growing number of published journals, conference proceedings, medical sites and portals on the World Wide Web, electronic medical records, etc. In particular, in the clinical context, there has been an increase of interest in the use of Electronic Medical Records *(EMR)* systems which may contain large amounts of text data, to improve the quality of healthcare [2]. To make full use of the information contained in the EMR and to support clinical decision, text mining techniques based on Natural Language Processing *(NLP)* have been especially proposed for information retrieval purposes or for extracting clinical summaries. In [3], an information extraction system that extracts three types of information (numeric values, medical terms and categories) from semi-structured patient records, is presented. An extension to this system is presented in [4]: The MEDical Information Extraction (MedIE) system extracts a variety of information from free-text clinical records of patients with breast related diseases. MedIE uses GATE [5], WordNet [6] and UMLS, and employs a graph-based approach for numeric attribute extraction capable of performing the majority of information extraction tasks achieving remarkable results. In [7], the

Keyphrase Identification Program *(KIP)* is proposed, for identifying medical concepts in medical documents. KIP combines two functions: noun phrase extraction and keyphrase identification. It automatically extracts phrases containing nouns using a part-of-speech tagger achieving fair results (0.26 in precision and 0.60 in recall, best case scenario). KIP ranks all the noun phrases in terms of their relevance to the main subject of the document, and selects only the most relevant ones by creating a glossary database from the Medical Subject Headings (MeSH) site. In [8] is presented a pipeline-based system for automated annotation of surgical pathology reports with UMLS terms built on GATE. The system implements a simple method for detecting and annotating UMLS concepts as well as annotating negations based on the NegEx algorithm [9], achieving very good results in terms of precision (0.84) and recall (0.80).

While all of these tools offer great insight on how concept identification and annotation can be done they do not offer any functionalities for single-document text summarization. Such feature can be found in more complex works, as in [10, 11] where summarization of single documents is done by applying robust NLP techniques combined with conceptual mapping based on ad-hoc ontologies or lexicons. The main problem with these approaches is that the accuracy of the concept extraction, and subsequently the accuracy of the summarization, depends on the underlying lexicon, and in this particular case, the ontology. Not using well established ontologies carries the drawback of limiting the available identifiable concepts and also, their interoperability with other semantic web-based complementary systems. In [12], *UMLS* is used for concept mapping but the system does not deal with negative expressions leading to misinterpretations in the fi-

nal summary.
In the next section the description of a system aiming at creating summaries out of medical records written in free text form by implementing a GATE pipeline, and also for assigning UMLS codes to the medical entities found inside them, is proposed.

2.3 Method

In order to produce a reliable summary, the corpus of medical documents must undergo through several processing steps. In this section, the tools used during this process are introduced and described. The basis of the developed system is GATE, which is the most used tool for implementing NLP-based applications. GATE uses regular expressions to configure all of its components (Tokenization, Sentence Splitter, POS tagging, Named Entity Recognition (NER) etc...).
The general architecture of the proposed system is shown in Fig. 2.1.

2.3.1 Text processing and annotation

ANNIE [13] is the information extraction component of the GATE platform and it substantially encapsulates the main NLP functions. In our case, an ANNIE pipeline was defined that employs the following components:

- **English Tokenizer**: The text in the corpus is divided into very simple tokens such as numbers, punctuation symbols or simple words. The main objective of this module is to maximize the efficiency and flexibility of the whole process by reducing the complexity introduced by the grammar rules.

- **Gazetteer**: Its role is to identify the names of entities based on lists, fed into the system in the form of plain text files. Each list is a collection of names, such as names of cities, organizations, days of the week, etc...

- **Sentence Splitter**: As its name suggests, it splits the text in simple sentences by using a list of abbreviations to distinguish sentence markers.

- **Part-of-speech Tagger**: Marks a word as corresponding to a particular part of speech based on both its definition and context. This is useful for the identification of words as nouns, verbs, adjectives, adverbs, etc. The results of this plug-in are the tokens used for the implementation of regular expressions.

- **Named Entity Transducer**: ANNIE's semantic tagger contains rules that work on the annotations of the previous phases to produce new annotations. It is used to create annotations regarding the terms related on negations, sections and phrases.

- **MetaMap Annotator**: This module serves the role of identifying medical terms found in text and map them to UMLS concepts by using NLP methods combined with computational linguistics [14].

- **Words Correction**: Given that the vast majority of the medical reports that we are dealing with were produced in a completely manual manner, misspellings do occur, making the medical term identification process less accurate. For this reason, each unannotated term (i.e. a word that does not exist) in the

text is used as a query term against a dataset containing medical terms and the term with the smallest Levenshtein distance is retrieved. The result is used in place of the misspelled word in the original document.

- **Negated Expressions**: In order to achieve a correct interpretation of the text found in medical documents, it is very important be able to identify negated expressions, which indicate the absence of a particular symptom or condition. MetaMap helps to identify negated concepts by providing a pair of features, namely "NegExType" and "NegExTrigger"; the former one identifies the negation, while the latter one specifies the term that expresses it. In this phase there are two problems that must be dealt with: a) the negated medical concept must be correlated to the term that triggers the negation effect and b) there are words that imply negation but MetaMap cannot identify them as such (e.g. the word inexistence). To overcome these problems, the Gazetteer is used again, by creating a new class of annotations relating exclusively to terms of negation.

- **Section parsing**: For this phase, the Gazetteer plug-in is used by defining tags that could be possibly represent section labels. For our experiments the following tags were defined: admitting diagnosis, discharge diagnosis, symptoms, past medical history, family history, social history, hospital course, medications, diagnostic studies, discharge instructions.

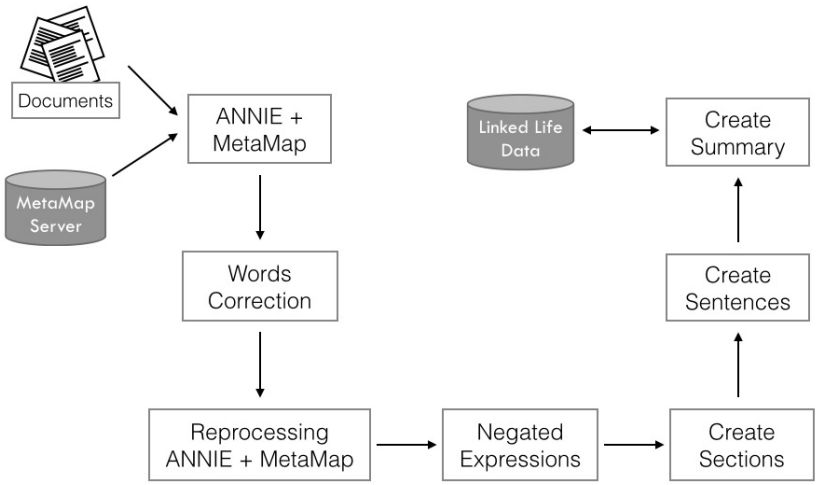

Figure 2.1: *General architecture of the proposed system.*

2.3.2 Summary Generation

Not all of the annotations generated by the MetaMap Annotator are needed in the final summary. Each MetaMap annotation contains also the semantic type of the corresponding term (e.g. "Body Part" for the word "leg", "Manufactured Object" for the word "scalpel" etc...). Inevitably, terms belonging to certain semantic types are excluded from the summary because their importance might be negligible.

An issue that needs to be dealt with during summary generation is that many annotated phrases should be merged to one sentence. For example, the sentence "x-rays including left foot, right knee, left shoulder and cervical spine" would normally be divided in the tokens "x-rays", "left foot", "right knee", "cervical spine" and "left shoulder"

even though all of them belong to the same sentence.

Regular expressions were employed to face this problem. In our case, the following regular expression was used:

$$(PRE)?(NEG)?((METAMAP)(NEG)?)+(POSTCONCEPT)?(POST)?,$$

where $METAMAP$ denotes the main medical concept identified by MetaMap (e.g. "amoxicillin", PRE denotes attributes that can precede the main concept (e.g. "significant", "treated with", "diagnosis of", "presence of" etc...), $POSTCONCEPT$ indicates a word directly correlated to the main concept (e.g. "1 g" for expressing dosage etc...) and $POST$ denotes eventual tokens that may represent a continuation of the sentence (e.g. commas, conjunctions etc...). Finally, the NEG term indicates whether a token expresses negativity or not.

The "+" and "?" operators describe the cardinality of each term with the "+" operator meaning "at least one or more" and the "?" operator meaning "zero or more".

For each identified section, the annotations relative to affirmative and negative expressions are created and for each sentence, the annotations produced by MetaMap are used. The same annotations are also used as query terms on the LLD site and the URLs pointing to the corresponding medical concepts are embedded to the final summary and exported in an HTML file.

An example of how the system works is shown below. Given the following discharge summary (the underlined words represent typographical errors):

ADMITTING DIAGNOSES: Intrauterine pregnancy at 36 weeks. Twin gestation. Breech presentation of twin A.
DISCHARGE DIAGNOSES: Intrauterine prengancy at 36 weeks. Twin gestation. Breech presentation of twin A. Status post primary low transverse cesarean section for malpresentation of twins.
CHIEF COMPLAINT: At the time of admission, contractions.
HISTORY: The patient is a 32-year-old pregnant at 36 weeks with known twins with contractions and good fetal movement, no bleeding, no loss of fluids.
OB HISTORY: Present pregnancy with previous receipt of a steroid window.
GYN HISTORY: Significant for chamydia, which was treated.
MEDICATIONS: Prenatal vitamins.
SOCIAL HISTORY: No drinking, smoking or drug use. No domestic violence. The father of the baby is currently involved, and the patient is living with a friend.
PHYSICAL EXAMINATION: Temperature is 36.2, pulse 88, respirations 18 and blood pressure 121/58. HEART: Regular rate and rhythm. LUNGS: Clear. ABDOMEN: Soft and gravid.
HOSPITAL COURSE: Postoperatively, the patient did well. She was eating, ambulating and voiding, passing gas by postoperative day 2, and on postoperative day 3, she continued to do well. She had been seen by Social Work and options made aware to the patient. She was ready for discharge. She remained afebrile throughout her hospital course.
DISCHARGE INSTRUCTIONS: She will be discharged to home to follow up in two weeks for a wound check.
MEDICATIONS AT THE TIME OF DISCHARGE: Percocet, Motrin and Colace.

The result is a more compact form of the input document, with both the wrong words corrected and also contains the Linked Life Data links identified by MetaMap:

ADMITTING DIAGNOSIS: Intrauterine pregnancy. Breech presentation of twin.
SYMPTOMS: contractions.
DISCHARGE DIAGNOSIS: Intrauterine pregnancy. Breech presentation of twin. Malpresentation of twins.
DIAGNOSTIC STUDIES: Temperature 36.2, pulse 88, respirations 18 and blood pressure 121/58. HEART. LUNGS. ABDOMEN.VAGINAL
PAST MEDICAL HISTORY : Significant for chlamydia. known twins with contractions and good fetal movement ,. pregnancy. Receipt of a steroid window.
PAST MEDICAL HISTORY NEGATIVE: no bleeding, no loss of fluids.
SOCIAL HISTORY NEGATIVE : No drinking, smoking or drug use. No domestic violence.
MEDICATIONS : Prenatal vitamins. Percocet, Motrin and Colace.

By clicking on the underlined terms, the system redirects the reader to its *LLD* page (Fig. 2.2).

2.4 Experimental Results

As stated in [1], evaluating the performance of a summarization system is not a trivial task. To be more precise, while the quantitative evaluation can be based on clear and objective metrics, the qualitative one is not that straightforward because summarization efficiency is most often expressed as a subjective opinion of the individual rater (i.e. Inter-rater reliability). Nevertheless, because of the two-fold nature of these kind of systems, their performance evaluation should cover both these aspects. So, in order to assess exhaustively the performance of the proposed system we tested it under three different perspectives and compared the results to a hand-crafted ground-truth

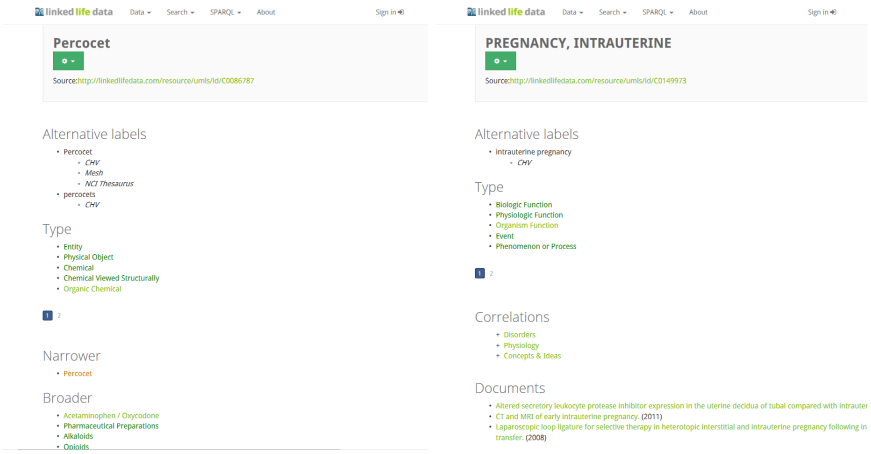

Figure 2.2: *Image showing the LLD pages of the terms Percocet (left) and Intrauterine pregnancy (right)*

(described in Subsection 4.1). For all the evaluations we employed Precision-Recall and F_1 measure values defined as follows:

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

,

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

and

$$F_1 = \frac{Precision \times Recall}{Precision + Recall}$$

The FP, TP and FN values are defined separately for each of the aspects tested. The obtained results were compared against a manually created dataset by five medical doctors that contained both positive and negative sentences. The dataset was comprised by 125 medical reports containing 3611 annotated sentences (2824 positive

and 787 negative) and 15641 annotated medical concepts.

- **Medical concept recognition**: The first aspect of the system that was tested was its ability to identify correctly the medical concepts found inside the medical reports.

  - A True Positive *(TP)* results when an identified medical concept is the same with the manual annotation.
  - A False Negative *(FN)* results when a medical concept was not identified correctly or was not identified at all.
  - A False Positive *(FP)* results when a medical concept was assigned a different label or when a non medical term was identified as such.

Test	N	TP	FP	FN	P	R	F_1
Medical concept recognition	15641	12499	2419	3142	0.84	0.8	0.82
Sentence identification and polarity detection	3611	2808	531	803	0.84	0.78	0.81
Medical concept recognition	15641	11499	3514	4142	0.77	0.74	0.75

Table 2.1: *Performance of the system in recognizing correctly the medical concepts.*

N	TP	FP	FN	P	R	F_1
3611	2808	531	803	0.84	0.78	0.81

Table 2.2: *Performance of the system on sentence detection and polarity detection.*

- **Sentence identification and polarity detection**: The second aspect of the system that was tested was its ability to extract correctly the single sentences in the medical report and also to assign correctly the negation attribute to the medical concepts detected by the previous test, using regular expressions.

  - A True Positive *(TP)* results when an identified sentence is found also in the ground truth and was assigned the correct polarity.
  - A False Negative *(FN)* results when a sentence found in the ground truth was not identified as such or when an annotated sentence was divided erroneously between two other sentences or when the negation property was not assigned to a negative sentence .
  - A False Positive *(FP)* when a sentence is erroneously identified as such, but instead, in the ground truth, its terms do not belong in the same one or when the negation property was assigned to a positive sentence.

- **Summary relevance**: Additionally, the quality of the produced summary was evaluated. To achieve this, the same five medical doctors were presented with both the original reports and the final results and then asked to assess qualitatively the relevance of the summaries (i.e. express their personal opinions on what medical concepts should be included in the final summary versus what should be excluded). After that, the following parameters were defined:

  - A True Positive *(TP)*: A concept that the medical doctors

N	TP	FP	FN	P	R	F_1
15641	11499	3514	4142	0.77	0.74	0.75

Table 2.3: *Performance of the system on summary accuracy. The final result was calculated based on the sum of the votes of the medical doctors.*

felt that should be included in the final summary and it was.

- A False Negative *(FN)*: A concept that the medical doctors felt that should be included in the final summary but it was not.
- A False Positive *(FP)*: A concept that the medical doctors felt that should not be included in the final summary but it was.

Sentence identification and polarity detection performance was very good. Indeed, an $F_1 score$ value of 0.81 means that the algorithms employed to do this task performed very well. More detailed inspection of the failing sentences were due to misplaced punctuation marks and missing negative keywords from the employed dictionary that could provoke ambiguity problems if they were ultimately included (e.g. the word *"will"* in the sentence *"...will develop cancer..."* does not imply that the patient has cancer). The results in medical concept recognition are almost equal as high. An $F_1 score$ value of 0.82 means that the MetaMap module is very accurate in identifying the medical concepts found in the reports. Especially important are the

results in the summary accuracy test where the subjective opinion of the intended end users of the system (the medical doctors) determine its utility, achieving an $F_1 score$ of 0.75.

CHAPTER

THREE

TIMES SERIES ANALYSIS: TRANSCRANIAL MAGNETIC STIMULATION

Transcranial magnetic stimulation (TMS) is the most important technique currently available to study cortical excitability. Additionally, TMS can be used for therapeutic and rehabilitation purposes, replacing the more painful and invasive transcranial electric stimulation (TES). In this chapter we present an innovative and easy-to-use tool that enables neuroscientists to design, carry out and analyze scientific studies based on TMS experiments for both diagnostic and research purposes, assisting them not only in the practicalities of administering the TMS but also in each step of the entire study's workflow. One important aspect of this tool is that it allows neuroscientists to specify research designs at will, enabling them to define any parameter of a TMS study starting from data acquisition and sample group definition to automated statistical data analysis and RDF data storage. It

also supports the diagnosing process by using on-line support vector machines able to learn incrementally from the diseases instances that are continuously added into the system. The proposed system is a neuroscientist-centred tool where the protocols being followed in TMS studies are made explicit, leaving to the users flexibility in exploring and sharing the results, and providing assistance in managing the complexity of the final diagnosis. This type of tool can make the results of medical experiments more easily exploitable, thus accelerating scientific progress.

3.1 Introduction

Transcranial magnetic stimulation (TMS) is a noninvasive and painless technique for the evaluation of corticospinal tract function as well as of motor cortex excitability of the human brain and it is used to investigate the central motor pathways of several neurological and psychiatric diseases. More specifically, TMS is the most important technique currently available to study cortical excitability [15], and can be used for therapeutic and rehabilitation purposes [16] and [17], replacing the more painful transcranial electric stimulation (TES). In the last twenty years, TMS has been applied to explore the pathophysiology of many neurological and psychiatric diseases [18], such as multiple sclerosis [19], stroke [20], dementia [21], Parkinson's disease [22], myelopathies [23], depression [24], schizophrenia [25], and as a possible therapeutic tool for some of these disorders [23].

TMS produces a modification of the neuronal activity of the primary motor cortex stimulated by the variable magnetic field generated

by a coil placed on the scalp. This variable magnetic field, produced by the current flowing in the coil, induces an electric current in the underlying brain tissue. The figure-of-eight or butterfly coil can stimulate a relatively focal area (Fig. 3.1), whereas the circular coil a more diffuse one [26].

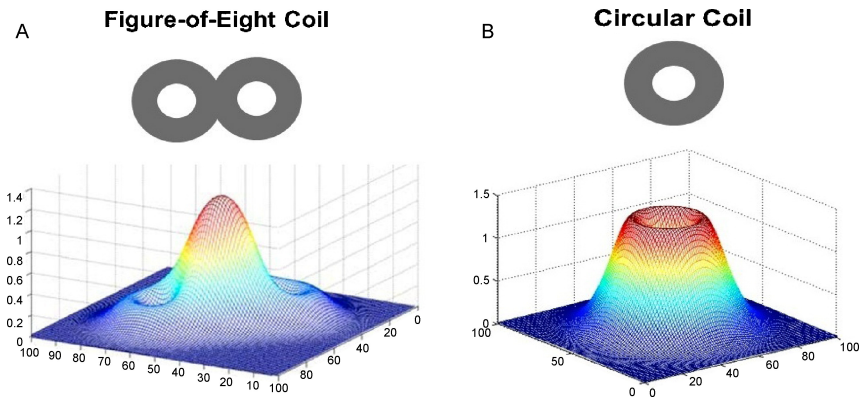

Figure 3.1: *Magnetic field generated by the different coils: (a) magnetic field by a figure-of-eight coil and (b) magnetic field by a circular coil.*

When TMS is applied to the primary motor cortex, at appropriate magnetic field intensity, it induces motor evoked potentials (MEP), recorded with an electromyograph, in the muscles that are contralateral to the stimulated motor cortex [27].

In clinical practice, TMS may be delivered as either single or paired pulses or regularly repeating pulses (repetitive TMS) in order to assess different parameters about the motor system. The single pulse TMS is used to evaluate the integrity of motor pathways and motor cortex excitability by measuring:

1. the MEP amplitude (defined as the distance between the lowest negative peak and the highest positive peak and expressed in mV);

2. the motor threshold (defined as the minimum TMS intensity necessary to evoke small-amplitude MEPs, larger than 50 V in amplitude);

3. the central motor conduction time (i.e. the latency difference between the MEPs induced by stimulation of the motor cortex and those evoked by spinal stimulation);

4. the cortical silent period (cSP, defined as a period of electromyographic suppression after a MEP).

Usually, the cortical excitability and intracortical circuits in various diseases are studied by a paired pulse TMS paradigm that couples a subthreshold stimulus (the amplitude is set lower than the patient motor threshold and it is called a conditioning pulse) and a suprathreshold stimulus (called a test pulse), at different interstimulus intervals (ISIs) through the same coil. The effects of the conditioning pulse on the size of the MEP depend on the duration of the ISIs. Indeed, at ISIs within the range 14 ms there is a strong inhibitory effect on the MEP (in the form of a reduced amplitude) [28], while at ISIs within the range 720 ms there is a facilitatory effect on the MEP (in the form of increased amplitude) [29].

Since there is an extensive use of TMS in different research fields and for each use of TMS several different factors are crucial, a data acquisition and processing system is required to create more standardized conditions and to reduce the high intra- and inter-rater variability

in the execution of the clinical experiments (typically due to coil positioning and to the time interval between each pulse administration).

As far as we know, very few software-based approaches have been proposed for supporting neuroscientists in performing TMS experiments. The first attempt was developed in 2000 by Kaelin-Lang and Cohen [30] who tried to help neuroscientists in the execution of TMS experiments, but the system was designed only for data acquisition and for data post-processing, and not for supporting researchers in the whole life-cycle of a research study. In order to improve the functionalities of this system, we have recently proposed a flexible TMS data acquisition and processing system affording the scientists an easy and customizable interaction with the TMS hardware, for more efficient and accurate data recording and analysis [31]. In this chapter we expand this work by presenting a system that, beyond the customization of the TMS experiments, uses machine learning techniques to assist scientists in the diagnosing process. In detail, here we propose an easy-to-use tool that enables neuroscientists to design, carry out and analyze scientific studies based on TMS experiments for both diagnostic and research purposes, and assists neuroscientists in each step of the entire study's workflow. The tool allows neuroscientists to specify any research design, by defining any parameter of a TMS study starting from data acquisition to sample group definition to statistical data analysis. All the data used in the proposed tool, including experiment protocol data, is also stored in RDF, thus they can be shared with other systems compliant to semantic web standards. Finally, the tool is also provided with on-line support vector machines (SVM) to help neuroscientists in the diagnosis process.

The remainder of the chapter is as follows: the next section intro-

duces the signals and the parameters involved in a TMS experiment. In Section 3 the proposed tool is presented, following each step of the workflow carried out by scientists for TMS experiments, from hardware interfacing to protocol definition, to experiment execution, to statistical analysis and RDF data storage. In the same section, the proposed on-line SVM approach for supporting scientists in the diagnosis is described, pointing out its advantages.

3.2 Transcranial Magnetic Stimulation

As mentioned in Section 1, TMS may be administered as either single or paired pulses or regularly repeating pulses (repetitive TMS). Single and paired pulses TMS are used for diagnostic purposes in order to assess different parameters about the motor cortex excitability, whereas repetitive TMS is used for therapeutic purposes. Investigating the motor cortex excitability involves measuring MEP amplitudes, motor threshold and silent period by using the single pulse TMS and the intracortical inhibition (ICI) and facilitation (ICF) by using the paired pulses TMS. The single pulse TMS consists of administering a single pulse and of recording the electromyographic (EMG) response, whereas TMS paired pulses consists of the administration of two pulses (a conditioning one and a test one) with a certain delay, called Inter-Stimulus Interval ISI. Fig. 3.2a shows the MEP response when a paired pulse stimulus is administered to a patient.

In such signals it is possible to identify:

- The latency, which is the time interval between the instant when the stimulation is administered to the subject and the instant

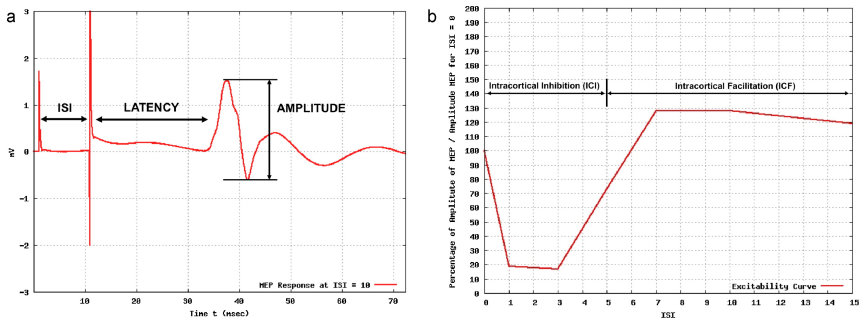

Figure 3.2: *Example of: (a) MEP response when a paired pulse TMS is administered to a patient and (b) cortical excitability curve.*

when the muscle starts to move. Latency tends to increase with age and height.

- The amplitude of the muscular response, which is the peak-to-peak excursion expressed in volts of the instrument that measures the muscle response.

The intracortical inhibition (ICI) and facilitation (ICF) are, instead, related to the cortical excitability that is estimated by a graph that describes the obtained amplitudes of the muscular responses at varying of the ISIs with respect to the amplitude obtained at ISI = 0. An example of a cortical excitability curve is shown in Fig. 3.2b. Currently all the TMS experiments are carried out by interacting manually with the TMS hardware, hence by setting only one ISI per time, whereas the number of repetitions for each ISI is performed by the experimenter by clicking a button on the coil as many times as the number of repetitions. Indeed, although the available TMS equipment is provided with tools allowing the automatic parameter setting, such

tools use proprietary script languages (similar to programming languages, e.g. the software Signal of the Cambridge Electronic Design) that make the task of designing TMS experiments very difficult and tedious for medical doctors.

In the next section the proposed customizable acquisition and processing system that permits the full customization of all currently used TMS paradigms (single pulse and paired pulse TMS) is described.

3.3 The Proposed Tool

This section describes a customizable data acquisition and processing tool that supports neuroscientists in the automatization and customization of all currently used TMS paradigms, in the data storage and experiment management and in the diagnosis. The architecture of the proposed system is shown in Fig. 3.3 and consists of three main modules:

- Hardware interaction module: it handles the interaction with the hardware equipment for executing TMS experiments;
- Experiment data management module: it allows neuroscientists, through an intuitive interface, to store patient data in RDF format, to set the parameters of TMS experiments, to process the acquired data, to define research studies involving several patients and to analyze data from such studies with statistical tests;
- Diagnosis support system module for supporting neuroscientists especially in the differential diagnosis. This module performs

on-line training from data to handle uncertain cases.

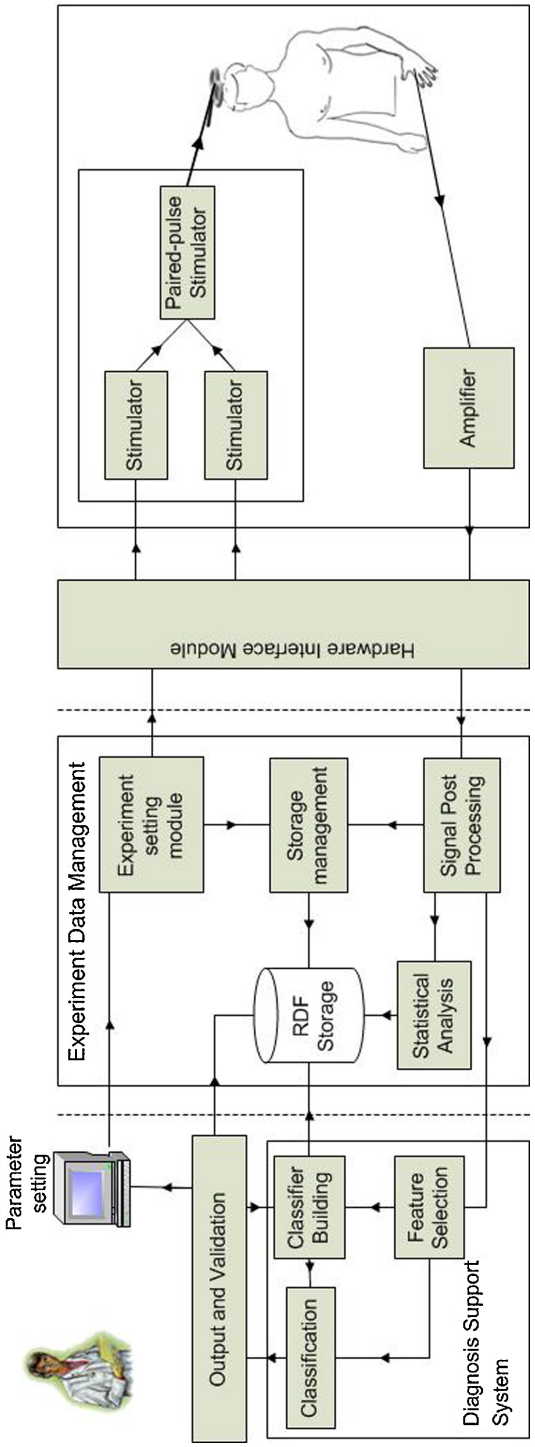

Figure 3.3: *The proposed architecture.*

3.3.1 Hardware Interaction Module

A hardware-interface communicates with the TMS equipment that interacts with a real-time data acquisition unit. This module implements a common programming interface in order to support different data acquisition systems. It is sufficient to import a library (specific for the hardware) for enabling the communication with the TMS hardware.

To date, only the library for communicating with the CED 1401[1] is present in our system. The CED 1401 A/D is one of the most common signal acquisition systems for TMS response acquisition and stimulation synchronizer and it usually comes with MagStim[2] stimulators. It features 4 analog inputs capable of acquiring signals with 16 bit resolution at a 500 kHz sampling rate, 2 digital inputs and 2 digital outputs. One of the analog inputs is used to acquire the response signals through a small signal amplifier (CED 1902). Therefore, the CED 1401 receives the user-commands, and synchronizes two stimulators MagStims 200, connected on its digital outputs, for the creation of the single pulses, which are further combined in a paired pulse by the Magstim BiStim and are administered to the patient's cortex through the coil. After the TMS stimulus administration, the muscular response (MEP) is registered by using single-use, low-noise, high conductivity electrodes. Such motor responses are then amplified, using the CED 1902, with a gain ranging from 100 to 1,000,000 (V/V) and a maximum voltage input range 10 V.

[1] http://ced.co.uk/

[2] http://www.magstim.com/

3.3.2 Experiment data management module

To assist the neuroscientists in the entire life-cycle of a TMS based research, the proposed tool provides the users with a set of flexible functionalities for setting all the necessary parameters, for processing the acquired data and for storing the information in order to be processed by other semantic-based applications or to be shared with other researchers. This module consists of four sub-modules:

- Experiment setting sub-module, for establishing the parameters of a TMS paradigm (ISI, number of repetitions, etc.), the criteria for patients enrollment and the variables (clinical, neuropsychological, etc.) of the patients that should be investigated for the specific scientific research;
- Signal post processing sub-module, for processing the acquired muscular responses in order to remove noise and other inconsistencies that may affect the quality of the acquired data;
- Statistical analysis sub-module, for assessing the results of the performed studies;
- Data storage sub-module, for handling the storage of any data produced in the system, from the patient's data, to statistical analysis results, to classifier's parameters. It is provided with different RDF repositories for each type of produced data.

Experiment setting module

Usually, a research study starts with the definition of a paired TMS protocol that involves the specification of the protocol variables to

be analyzed (clinical, psychiatric, neurophysiological, etc.) that are strictly related to the disease/diseases under investigation, and the TMS parameters, namely the ISIs to administer, the number of repetitions for each ISI and the modality of administration (random or sequential). The schema of this module is shown in Fig. 3.4 and the graphical user interface for protocol definition is shown in Fig. 3.5.

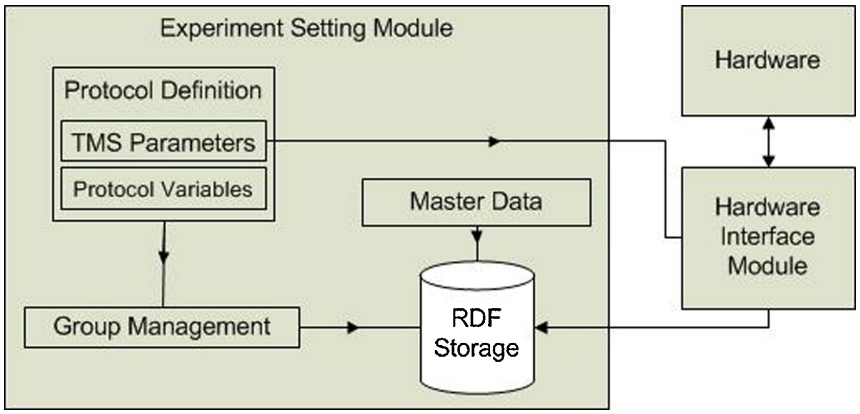

Figure 3.4: *Experiment data management module's architecture.*

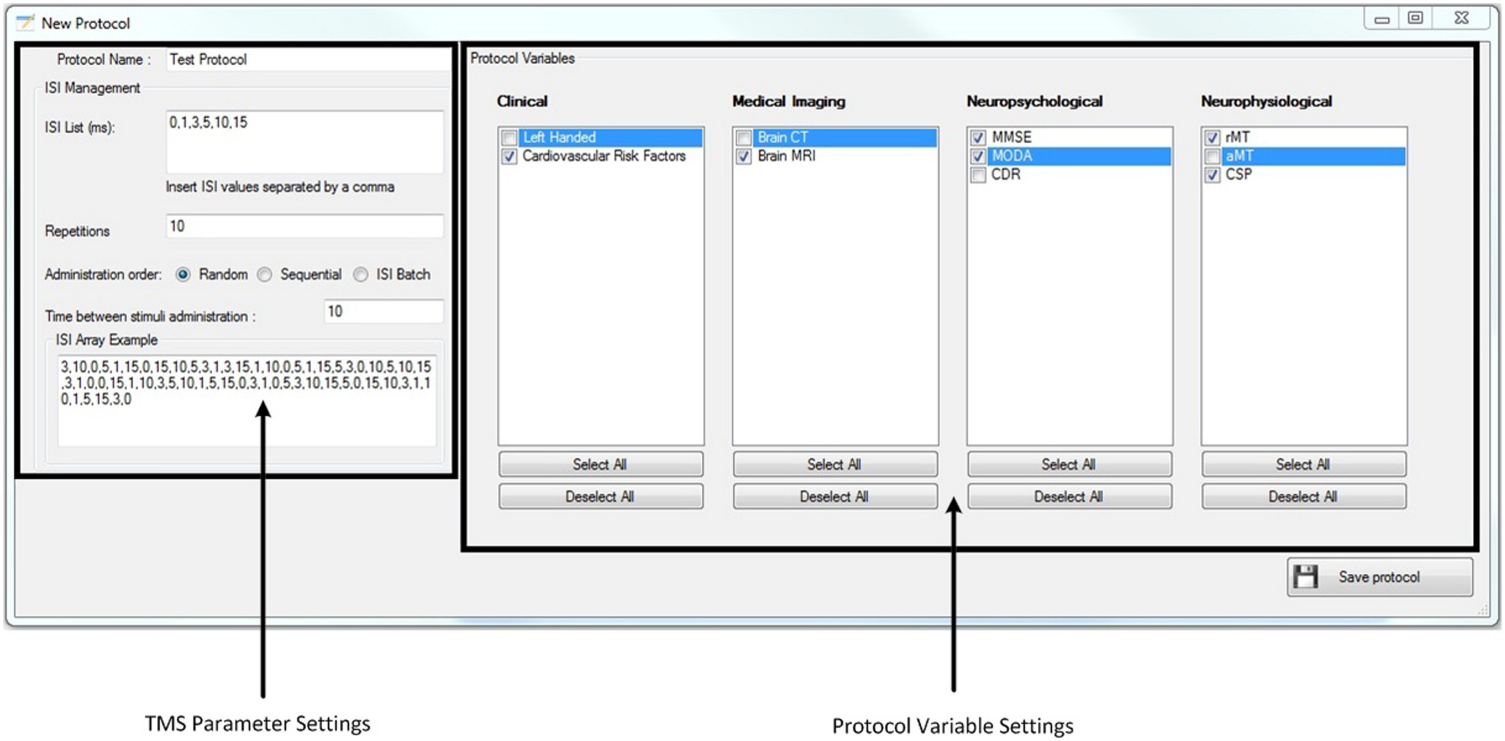

Figure 3.5: *The graphical user interface for protocol definition.*

After the protocol definition for a paired-pulse TMS, the data of each patient belonging to a specific study can be acquired. Among the variables specified in the protocol for each study, the common neurophysiological parameters such as motor threshold, silent period must be estimated using the single pulse TMS. After entering relevant demographic/clinical data of the patient under investigation, the paired-pulse TMS with the parameters set during the protocol definition can be administered to the patient. Fig. 3.6 shows the user interface while administrating paired-pulse TMS (with MEP responses) according to a specific protocol: in the left side the plots of MEP responses for a specific ISI are shown, whereas in the right side the TMS protocol settings are listed and the monitoring of the subject's relaxation status is displayed.

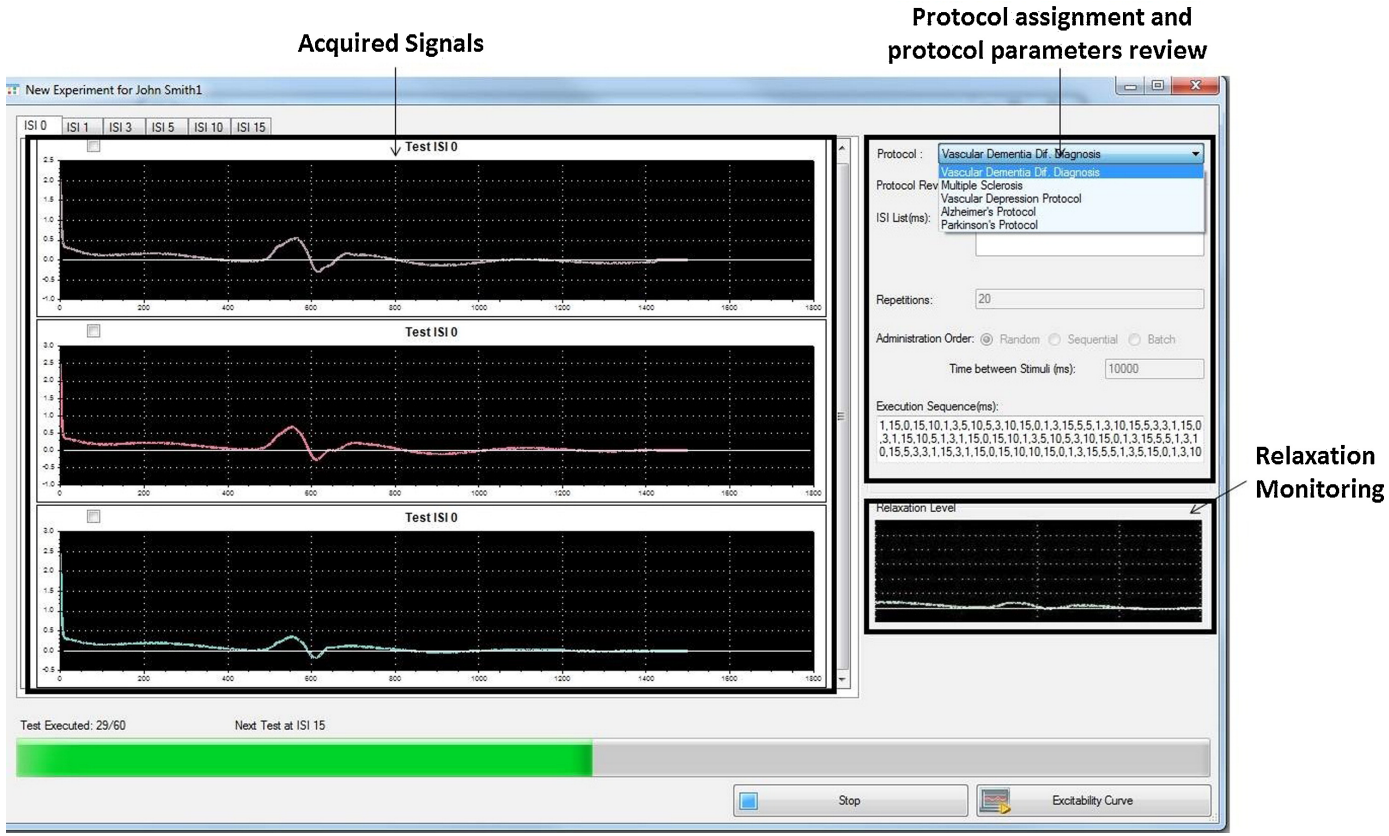

Figure 3.6: *The graphical user interface for protocol execution.*

Signal post processing module

After protocol setting and execution, the acquired data are processed by the signal post processing module. Indeed, the acquired MEP responses to the administered TMS stimuli rarely respect the quality criteria imposed by the experimenter because of both the variability of the MEP signals during the recording and the noise affecting such signals. MEP signals show, typically, a high variability in the values depending on two main factors: (1) the misalignment of the coil over the patient's head, which can be corrected by adjusting the coil's position, and (2) the stimulus administration when the relaxation level of the patient invalidates the muscular response; indeed, if the patient is relaxed the muscular response is generally accurate, whereas if the patient is nervous, suffers from a disease or is on medications that alter the electrical signals that the brain sends to the peripheral nerves, the acquisition of muscular responses is difficult, and often not possible. To deal with this problem, the proposed system includes an on-line monitoring module (right side in Fig. 3.6) that continuously evaluates the relaxation level of patients. This module checks the patient's relaxation level in real-time and eventually informs, in case of inappropriate levels, the experimenter, who can discard manually the acquired signals. Moreover, the system can be set to discard automatically the MEP responses according to the evaluated relaxation condition. The automatic MEP signal elimination is implemented by estimating if, at the time of the pulse administration, the relaxation level (computed as the area under the muscular response detected by the EMG, e.g. the curve of the MEP signal shown in the right side of Fig. 3.6) is in the range $\mu \pm \sigma$ where μ and σ are, respectively, the mean and the

standard deviation of the previously evaluated relaxation level.

The accuracy of the acquired signals may be also influenced by noise. For example, high amplitude $50 - 60$ Hz alternate currents are commonly found in any intrinsically noisy environment such as hospitals. The $50 - 60$ Hz AC noise is easily predicted and it can be removed by using notch filters in the appropriate frequency range ($49 - 51$ Hz for Europe, $59 - 61$ Hz for the USA). Another type of environmental noise is the high frequency interference due to the usage of other electrical/electronic devices near to the TMS acquisition equipment. Unfortunately, it is difficult to eliminate such noise without altering substantially the base response signal thus our system is provided with a noise removal tool based on Fourier Signal Decomposition. This tool addresses only sinusoidal and predictable noise by analyzing the signal's frequency components and therefore, the validation of the results is based on visual inspection carried out by the experimenters. The tool permits to re-administer a stimulus if the noise cannot be removed.

Statistical analysis module

After completing the data acquisition phase from the subjects sample, according to the designed protocol, the statistical analysis is performed. Usually, this step is done by a statistician, but often medical research centers are not provided by a statistics unit and this is a bottleneck. Therefore, the proposed system implements a statistics module that performs the most common tests for data statistical analysis, in an automatic and transparent way. This module exploits the functionalities of the IBM SPSS software by using the SpssClient API.

Depending on the variables that the study's protocol contains, the statistical analysis module is able to automatically decide the appropriate statistical tests to perform. Moreover, according to the distribution of the values of the variables involved in a defined protocol, a specific test is selected. For example, in Fig. 3.7 we have the summary of a TMS protocol carried out on two groups of patients: control patients and patients affected by vascular depression. The variables defined in the protocol are Mini Mental State Examination (MMSE), Familiar History (F-Hyst), Personal History (P-Hyst) and the average (averaged on the number of repetition of each ISI) amplitude at ISIs 1, 3, 5, 7, 10 and 15 extracted from the cortical excitability curve.

Statistical Analysis Input Set

Name	Surname	Protocol	Initial Diagnosis	Smoker	Sex	Age	Left Handed	MRI Lesion Load	MMSE	Hypertension	P-Hyst	F-Hyst
John1	Smith	Vascular Depress...	Vascular Depress...	True	M	64	False	4	13	True	False	True
John2	Smith	Vascular Depress...	Control	True	M	78	False	1	23	True	True	True
Joan3	Smith	Vascular Depress...	Control	True	F	69	False	1	20	True	False	False
Joan4	Smith	Vascular Depress...	Vascular Depress...	False	F	60	False	0	29	False	True	True
John5	Smith	Vascular Depress...	Vascular Depress...	True	M	72	False	5	7	False	True	True
John6	Smith	Vascular Depress...	Vascular Depress...	True	M	64	False	4	13	True	False	False
John7	Smith	Vascular Depress...	Vascular Depress...	True	M	75	False	2	19	True	False	False
Joan8	Smith	Vascular Depress...	Control	True	F	83	False	1	27	True	False	False
Joan9	Smith	Vascular Depress...	Control	False	F	64	False	1	26	False	False	True
John10	Smith	Vascular Depress...	Vascular Depress...	True	M	77	False	5	7	False	False	True

Figure 3.7: *A subset of the patients group on which the vascular depression protocol has been executed. The comparison between these two groups (controls, vascular depression) is performed automatically by means of statistical tests.*

According to the type of variable to be compared, our system checks if the variable is a numeric value or a percentage and also performs the normality test to decide if parametric or non-parametric tests should be executed. In the case shown in Fig. 3.7 we have that the MMSE is a numeric variable and it is not normally distributed, therefore the MannWhitney test is performed, whereas since the variable P-Hyst is boolean, the comparison between the two groups is performed using the Chi-square test. Fig. 3.8 shows the output of the statistical analysis for the above described example.

Mann-Whitney Test

Ranks

	Disease	N	Mean Rank	Sum of Ranks
MMSE	1,00	15	9,67	145,00
	2,00	10	18,00	180,00
	Total	25		

Test Statistics

	MMSE
Mann-Whitney U	25,000
Wilcoxon W	145,000
Z	-2,779
Sig. (2-tailed)	,005

disease * Physt Crosstabulation

			Physt		
			,00	1,00	Total
disease	C	Count	9	1	10
		Expected Count	8,3	1,7	10,0
		% within disease	90,0%	10,0%	100,0%
	VD	Count	10	5	15
		Expected Count	12,4	2,6	15,0
		% within disease	66,7%	33,3%	100,0%
	Total	Count	29	6	35
		Expected Count	29,0	6,0	35,0
		% within disease	82,9%	17,1%	100,0%

Chi-Square Tests P Hyst

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5,196[a]	2	,074
Likelihood Ratio	6,473	2	,039
N of Valid Cases	35		

Figure 3.8: *Results of the statistical tests performed on the patients group shown in Fig. 3.7*

Data storage module

The nature of the data processed by the proposed platform permits the adoption of semantic repositories to be used as the system's storage servers. In fact, by using well established ontologies, like FOAF, and controlled vocabularies, like MeSH[3], and by creating an appropriate schema to describe the whole data structure, including data relationships, data can be easily processed by intelligent medical systems, such as the one proposed herein, and by semantic tools. In particular, the whole experiment workflow is enriched with information following an RDF schema that includes:

- The FOAF ontology to describe patient and neuroscientists information;
- The MeSH controlled vocabulary for coding disease, symptoms and signs associated to diseases;
- A set of RDF classes and properties that describe a TMS based scientific study including protocols, variables and TMS technical parameters.

A complete description of the RDF schema is beyond the aim of this chapter, although we provide here some highlights about the underlying design. The variables used in the protocol definition are stored in RDF and are structured as a SKOS vocabulary. In detail, they are grouped in several categories and we have a SKOS collection for each category, e.g. for clinical variable, for neurophysiological variable, for neuropsychological variable, for medical imaging variable.

[3]https://www.nlm.nih.gov/mesh/

We have also defined a class TMSProtocolVariable, for describing the variables (different from the ones above listed) that can be derived only from the TMS, which is also a subclass of SKOS:concept. This allows us to create a collection of TMS variables and to add other features (such as the range of the variables) that are not included in SKOS. In Fig. 3.9 an example of an RDF instance of the proposed schema and describing a generic TMS study is shown.

Personal information about the patient is inserted exclusively by the neuroscientist who carries out the examination and, for privacy purposes, our semantic system replaces the patient's FOAF profile URI with an appropriate MD5 hash string. The data storage has been implemented by semantic repositories using SESAME servers (see Fig.3.10) to make these information available for other purposes. In detail, four distinct RDF repositories are available:

- The patient master data store is the semantic database where all the information about patients is stored, including parameters for statistical analysis, like age, smoker, gender, etc.

- The variables data store is used for the variables defined during the TMS protocol design.

- The experiment data store is a combination of a semantic repository and a file server. The semantic database stores signal information, such as amplitude, latency, ISI. The file server retains the whole muscular responses in order to extract the aforementioned values and to export the acquired signal in a human readable format (e.g. an image).

- The classifier data repository, where the classifier's parameters for diagnosis support are stored.

Listing 1: RDF instance describing a TMS study

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:tms="http://i3s-lab.unict.it/semweb/ns/tmsSchema"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <tms:TMSStudy rdf:about="http://i3s-lab.unict.it/semweb/TMSStudies/2011/03/17/
       VascularDepressionTestStudy">
    <dc:description>A test protocol for assessing Vascular Depression in 10 patients</dc:description>
    <tms:Neurophysician rdf:resource="http://i3s-lab.unict.it/semweb/foaf/John_Smith"/>
    <tms:TMSProtocol>
      <tms:TMSParameters>
        <tms:ISIList>0,1,3,5,7,10,15</tms:ISIList>
        <tms:IBS>10</tms:IBS>
        <tms:Repetitions >15</tms:Repetitions>
      </tms:TMSParameters>
    </tms:TMSProtocol>
    <tms:TMSProtocol>
      <tms:TMSVariableCollection>
        <skos:member rdf:resource="http://i3s-lab.unict.it/semweb/ns/VariableDictionary#Smoker"/>
        <skos:member rdf:resource="http://i3s-lab.unict.it/semweb/ns/VariableDictionary#LeftHanded"/>
        <skos:member rdf:resource="http://i3s-lab.unict.it/semweb/ns/VariableDictionary#aMT"/>
        <skos:member rdf:resource="http://i3s-lab.unict.it/semweb/ns/VariableDictionary#CSP"/>
        <skos:member rdf:resource="http://i3s-lab.unict.it/semweb/ns/VariableDictionary#
            BrainMRILesionLoad"/>
      </tms:TMSVariableCollection>
    </tms:TMSProtocol>
    <tms:TMSExperiment>
      <tms:PatientCollection>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            c4ca4238a0b923820dcc509a6f75849b"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            c81e728d9d4c2f636f067f89cc14862c"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            eccbc87e4b5ce2fe28308fd9f2a7baf3"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            a87ff679a2f3e71d9181a67b7542122c"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            e4da3b7fbbce2345d7772b0674a318d5"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/1679091
            c5a880faf6fb5e6087eb1b2dc"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/8
            f14e45fceea167a5a36dedd4bea2543"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            c9f0f895fb98ab9159f51fd0297e236d"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/45
            c48cce2e2d7fbdea1afc51c7c6ad26"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/
            d3d9446802a44259755d38e6d163e820"/>
        <tms:Patient rdf:resource="http://i3s-lab.unict.it/semweb/TMSStudies/Patients/6512
            bd43d9caa6e02c990b0a82652dca"/>
      </tms:PatientCollection>
    </tms:TMSExperiment>
  </tms:TMSStudy>
</rdf:RDF>
```

Figure 3.9: *Example RDF schema instance.*

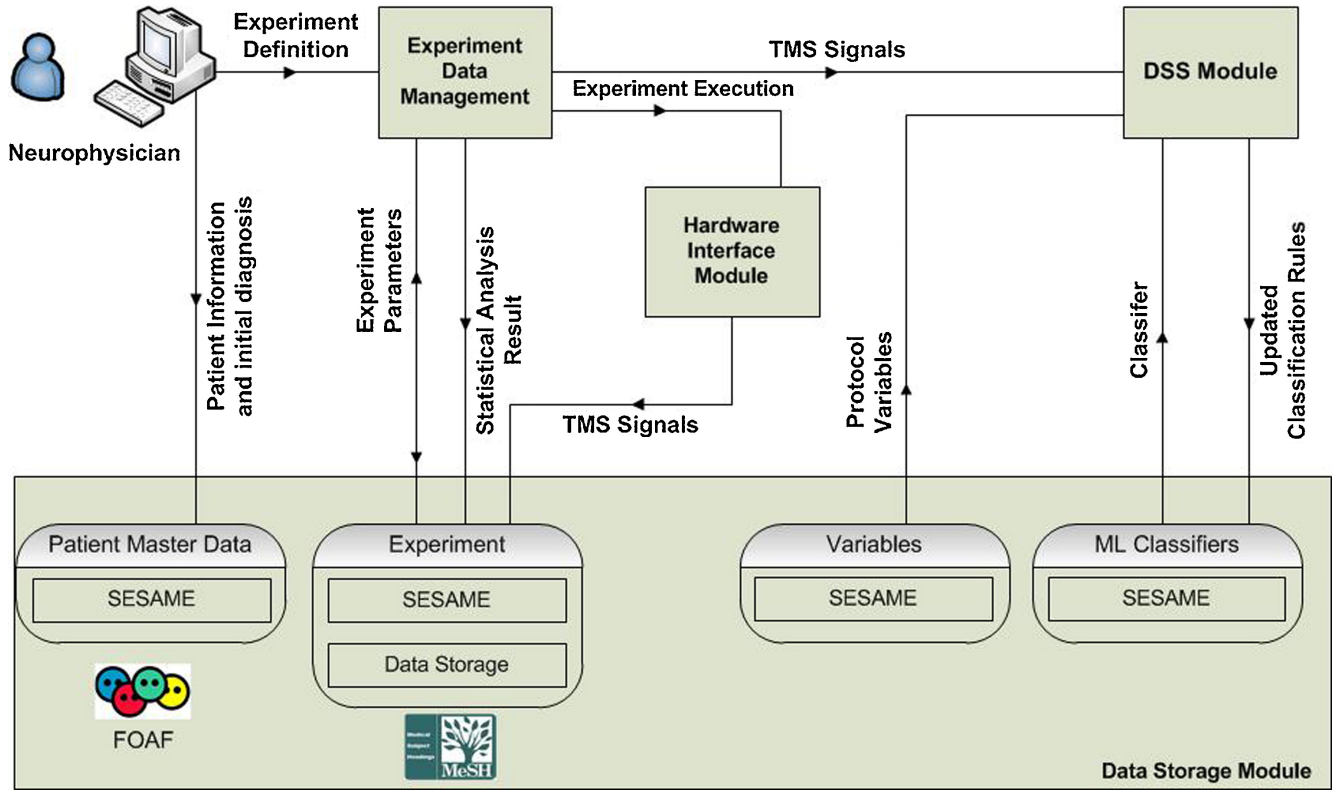

Figure 3.10: *Interaction between the data storage module and the other modules of the proposed system.*

3.3.3 Diagnosis support system

Currently, the diagnosis of many neurodegenerative and vascular diseases is mainly based on clinical evidence and on imaging techniques such as MRI, PET and SPECT. Often, especially at a very early stage, the clinical evidence of neurodegenerative disorders (e.g. Parkinson disease, Alzheimer disease, etc.) may be very similar to the one of vascular diseases. Medical imaging techniques (especially MRI) may help in such cases: indeed the MRI shows mainly atrophy of the brain in neurodegenerative disorders [32] and ischemic lesions in vascular diseases [33]. The problem arises when both types of evidence are present in an MRI, especially in elderly people who may have brain's atrophy due to the advanced age, although the main cause of their symptoms could be a vascular disease [34]. An example is the mixed dementia, i.e. the case when neurodegenerative dementia and vascular dementia occur at the same time [35]. The differential diagnosis is difficult not only in the above cases, but also among neurodegenerative diseases (e.g. Alzheimer disease vs Lewy body disease [36] or Parkinson disease vs Lewy body disease [37]) that could exhibit similar features at early stages. Therefore, it is necessary to identify the main cause of the observed signs and symptoms in order to provide the appropriate treatment. As mentioned in the introduction, TMS-studies have demonstrated, by investigating motor threshold, cortical silent period ICF and ICI, that the various neurological diseases may involve motor pathways in different ways. Hence, given that TMS provides detailed information about the motor system and since motor system's alterations have been identified in many neurological diseases, an appropriate processing of MEP responses may be used for

supporting the diagnosis.

Under this scenario, a diagnosis support system may play a double role: first, assess if the obtained MEP responses are evidence of neurological disorders, and second, support neuroscientists in differential diagnosis. To address the first need, two methods [38] and [39] have been proposed for classifying diseases such as Alzheimer and Subcortical ischemic vascular dementia by analyzing the MEP responses of a TMS paradigm. In particular a fuzzy system [39] and a neural network [38] were proposed and assessed for the differential diagnosis of Alzheimer and Vascular Dementia by using the following features: latency, amplitude, max and min module of the Fourier Transform, max and min module of the Hilbert transform of the MEP responses for ISI 1, 3, 5, 7, 10, and both of them achieved an average accuracy of about 92%. However, since these approaches are disease-specific (the training is done off-line) they cannot be used in a dynamic research and clinical context, such as the one here foreseen, where different TMS paradigms may be implemented for analyzing different diseases. For all of the above reasons, the proposed system is provided with an on-line diagnosis support system (DSS) that uses the above features extracted from a MEP response and it is based on a modified version of a support vector machine for large-scale problems (typically, about 1000 exams per year are executed in a single neurophysiological unit), capable of learning incrementally (averagely, between three and five exams per day are executed). Support vector machines (SVM) have been widely used for implementing classifiers because of their good generalization property [40]. Their main shortcoming is that training is time consuming, thus preventing their use in large-scale problems such as the one at hand. A solution is to resort to a modified SVM

that supports on-line incremental learning. Several approaches for incremental learning have been proposed. The first attempts were developed by Syed et al. in [41] and by Ruping in [42] by re-training the SVM through new examples combined with the already computed support vectors; however, these approaches are very memory demanding. Differently, to address large-scale issues, approaches based on clustering techniques for down-sampling the size of the examples and using the most representative ones for re-training have been proposed [43] and [44]. Therefore, the problems to be solved for on-line SVM are: the on-line selection of the learning data and the re-use of the already computed support vectors. Our diagnosis support system relies on the on-line SVM proposed in [45] that implements on-line training and, at the same time, solves the large scale problem. A detailed evaluation of the achieved performance, in terms of accuracy, and the comparison with the existing on-line classification systems are beyond the aim of this article, although we can report that in 18 uncertain diagnosis cases, over a totality of about 70 patients, the DSS performed well identifying the four diseases these cases belonged to. The DSS module is, therefore, used when a new patient whose diagnosis is unknown is inserted into the system (see Fig. 3.11 for the related GUI).

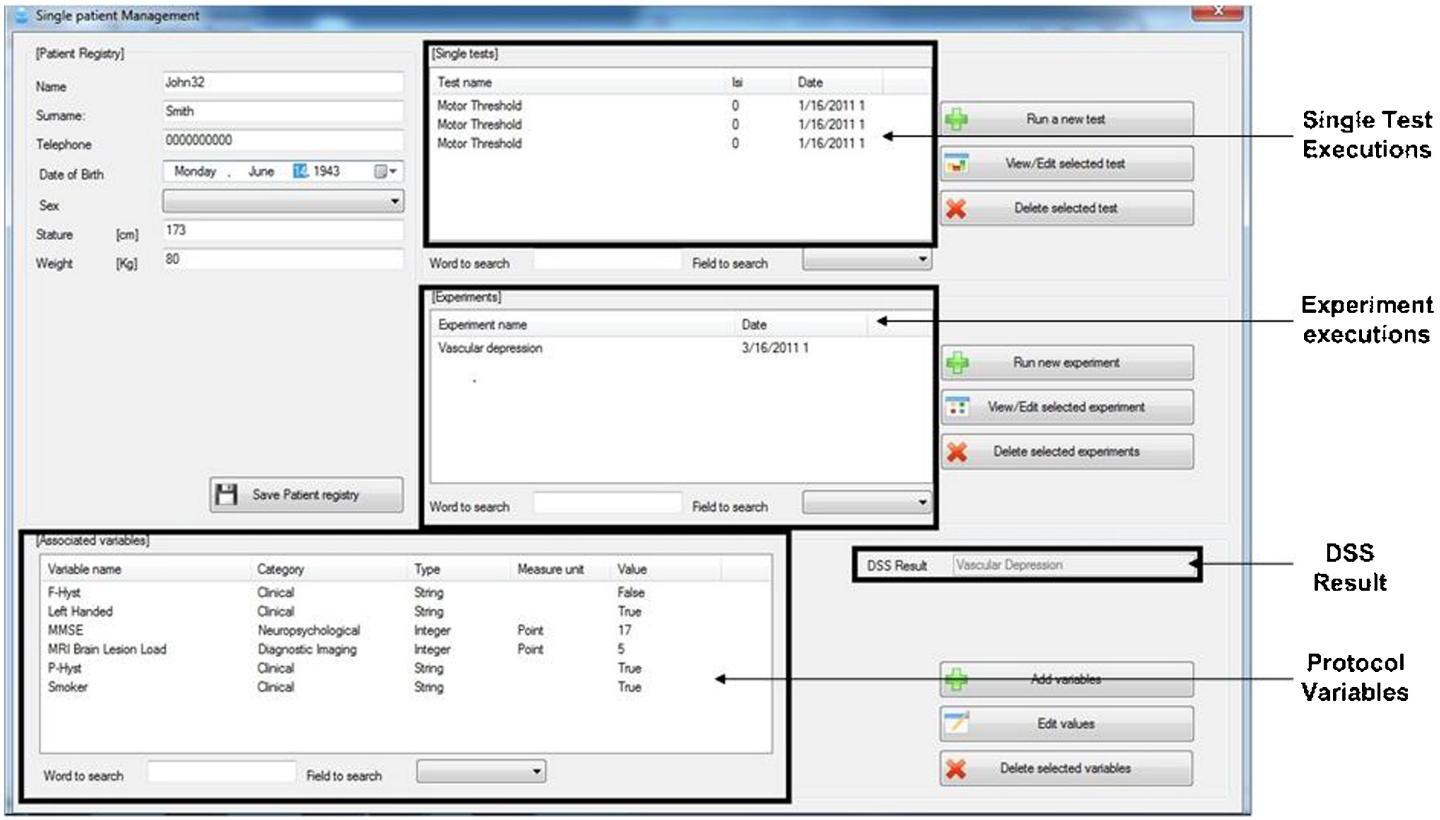

Figure 3.11: *Interaction between the data storage module and the other modules of the proposed system.*

CHAPTER

FOUR

IMAGE PROCESSING: SKELETAL BONE AGE MODELING BY HIDDEN MARKOV MODELS

Having an objective means to evaluate accurately the effective age of individuals, is a problem far from been resolved. Such solution would be very useful in many contexts: from pediatrics, to criminal investigation and to human rights. Assessing effectively the skeletal bone age based on X-Ray images is one way to achieve it but, given the excessive variability in the of the human species machine learning methods are employed to tackle the problem of universal application.

There are two globally recognized methods for bone age assessment: the Greulich and Pyle method (GP), which is based on the comparison of the X-Ray with an atlas, and the Tanner and Whitehouse method (TW2), which compares the developmental state of a set of bones.

In this chapter, a method and a tool for assessing the skeletal bone

age using X-Ray images of hands, implementing the TW2 bone age assessment method, is presented. The method combines image processing methods for enhancing the quality of the input images and Hidden Markov models for the classification task. The method was tested on a dataset made by two expert radiologists and its performance compared against state-of-the-art methods achieving very high accuracy in the evaluation of the skeletal bone age.

4.1 Introduction

The advancements in computer science have always boosted a large number of scientific fields by both facilitating and hastening the execution of repetitive and/or complex tasks. Image processing methods, in particular, have been used in a variety of applications in diagnostic medicine since their mere conception [46, 47] improving the diagnostic accuracy [48, 49].

Projection radiography was the first non-invasive method to depict the internal structures of the human body and it is currently one of the most used imaging methods. During the last decades a notable increase of interest in determining accurately the bone age by processing X-Rays, has been observed. This interest arises from the fact that having an accurate and, more importantly, objective assessment of the age of an individual results useful in many applications ranging from detecting and evaluating hereditary, hormonal or developmental disorders [50, 51] to creating indisputable evidence in legal cases where the real age of a person can determine his eligibility for criminal sanctions [52], legal rights [53] etc.

In the clinical practices two different approaches have been used for skeletal bone age assessment: the Greulich and Pyle (GP) [54] and the Tanner and Whitehouse (TW) [55] methods. Both approaches have been tested from the scientific community and their validity is already confirmed [56, 57, 58, 59].

The GP method, which is the simplest and most intuitive one, relies on comparing a subject's X-Ray of the left wrist to a gold standard atlas categorized according to age and sex. The TW2 method uses a-priori knowledge and creates a detailed analysis of the features of twenty predetermined regions of interest (ROIs) located in the left hand's bones, including epiphysis/metaphysis ROI (EMROI), carpal ROI (CROI), radius, and ulna. Each ROI is evaluated by assigning to it a letter, which represents the developmental status, ranging from A, meaning that the bone is completely absent, to I, which represents a fully developed, mature bone. As a final step, by summing up all the ROI scores the effective bone age is calculated.

The GP method is less complicated and generally faster to implement than the TW2 method. However, the latter offers better reproducibility and accuracy [60] and, because of its modular nature, TW2 is proner to automatization [61].

Although much research has been carried out the problem of estimating accurately the bone age of an individual, is far from being solved. This is demonstrated by the evergrowing number of surveys and future directions works (e.g. [62, 63, 64]).

In this chapter, we present a method and a tool aiming at determining skeletal bone age based on X-Rays of the left wrist using a modified version of the TW method based on EMROIs only, combined with Hidden Markov Model-based classifiers for refining the obtained

results. In the next section a review of the existing approaches is found while Section 3 describes in detail the application and its inner workings. The last section shows some performance measurements obtained by the actual usage of the system.

4.2 Related Works

Early attempts for "automating" the process of skeletal bone age assessment can be dated back to the early 90's, and in particular, in [65] where the authors present the first system employing simple image processing techniques, namely Sobel Gradient and thresholding, in order to make the image more suitable for the bone age assessment task. Measurements of the phalanxes were compared to the standard phalangeal length table [66] and the effective age was calculated. While this method suffered from the classical "infancy" problems (e.g. image quality, reproducibility etc..), it can be considered as one of the first steps towards more complex and accurate systems for bone age assessment. The methods for assessing skeletal bone age can be categorized in three main groups: fuzzy based, deformable models and machine learning mainly trying to reproduce the TW2 method.

Many methods have been proposed for dealing with the skeletal bone age estimation, by using fuzzy logic-based approaches. In [67] the authors present an automatic skeletal bone age assessment system for young children (from 0 to 7 years old) using only carpal bones. This method initially employs fully automatic carpal bone segmentation and morphological feature analysis and subsequently applies a fuzzy classification approach in order to assess the real bone age.

Other fuzzy-based methods combined with morphologic features of the carpal bones can be found in [68] where the authors also integrate Principal Component Analysis and statistical correlation or Support Vector Machines [69] in order to build a growth model of the carpal bones, declaring a success rate of 87%-89%, although they considered a relatively large admissible error of 1.5 years. Fuzzy classifiers are used in [70] for automating the GP method. Although it achieves a very high accuracy rate at lower age groups (0 to 2 years), its performance deteriorates when X-Rays of older subjects were used.

Deformable models (and especially Active Shape Models) have been largely used for skeletal bone age assessment [71, 72, 73]. Despite deformable model based approaches are capable of modeling EMROI shapes, they are ad-hoc solutions relying on many parameters with results that depend largely on the quality of the input images. The authors in [74] suggest that one of the main difficulties in assessing the age of an individual, is the irregular (i.e. largely varying) development of the trapezium and trapezoid bones and they propose a method, based on the integration of anatomical knowledge and trigonometry theory for the TW2 assessment.

Machine learning techniques have been also employed in automatic skeletal bone age assessment systems. In [75], the X-Ray image is segmented by using a K-means clustering algorithm applied on a gray-level co-occurrence matrix but, even though it is stated that the accuracy of the method is high, no performance evaluation was carried out extensively. A Support Vector Machine and correlation prototypes [76], and in [77], Support Vector Regression and smart class mapping have been proposed that, however, perform poorly in terms of accuracy.

Contrary to the majority of the existing systems based on a single evaluation method, BoneXpert [78] is a system for automatic skeletal bone age assessment that combines both the TW and GP methods. The main drawback of BoneXpert, however, is its high image rejection rate, meaning that it does not process low quality images and it often requires a heavy preprocessing step in order to make the image appropriate for processing.

While there exist many computer-based EMROI classification systems that employ machine learning approaches (e.g. Neural Networks, Fuzzy Classifiers, Support Vector Machines etc.) one of the main limitations is the lack of methods to model bone shapes effectively and dynamically. To deal with this issue we employ Hidden Markov Models which is a model of a sequential process changing states at discrete sequence intervals thus able to model ROIs' discrete stage. A further contribution of this chapter is the integration of several existing works, from preprocessing to finger extraction to stage assignment, into a unified tool which can be used by clinicians. In the following section we present a new approach, which extends our previous work [74], for efficiently assessing the skeletal bone age from an X-Ray of the left hand of a subject.

4.3 The Proposed Tool

Generally, the existing applications for skeletal bone age evaluation follow a standard workflow (Fig. 4.1) model. In such model, the input image is initially processed by noise removal (for enhancing the clearness of the input image) and background subtraction algorithms

(aiming at identifying the parts necessary for classification). Many algorithms, optionally include a machine learning step to aid in the classification process.

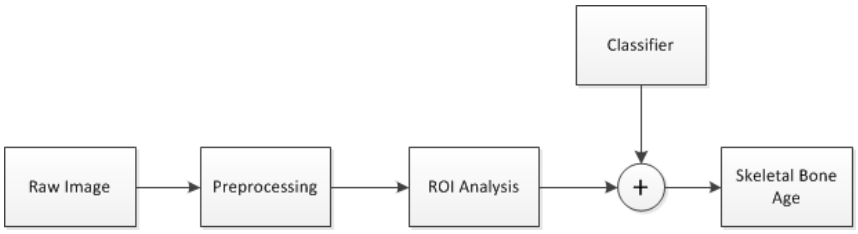

Figure 4.1: *Generic workflow of the currently existing skeletal bone age assessment tools. Note that the classifier stage is optional (i.e. it does not exist in all the systems).*

Before the presenting the method, a brief description of the TW2 method is necessary.

The Tanner-Whitehouse Method

The TW2 method is based on a predefined standard of bone maturity depending on age. It employs 20 ROIs located on the first, third and fifth finger and the carpus. The finger ROIs are called EMROIs (Epiphyses/Metaphyses ROI) and the carpal ones (including the long bones radius and ulna) CROIs (Fig. 4.2). The maturity of the bone is determined by the state of the epiphysis: if it is completely absent it represents an initial developmental state, else, if it is completely fused to the metaphysis bone maturation has completed. This development progression of each ROI can be divided into descrete stages, with each one assigned a letter from A (epiphysis is absent) to I (epiphysis-metaphysis fusion complete). A numerical score is further associated

with each stage of each bone and the overall maturity score is calculated by adding the individual ones. This score is then used to find the age according to the graphs in Fig. 4.3.

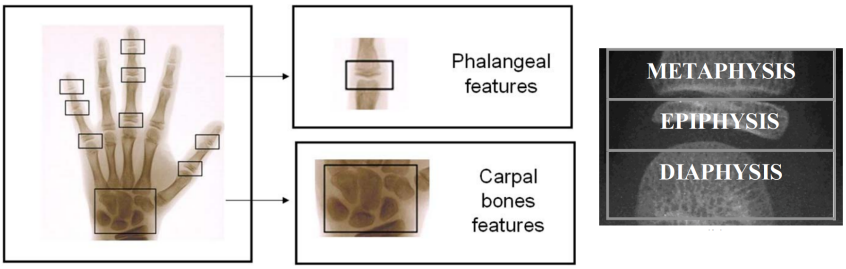

Figure 4.2: *The bones considered for the assessment of the bone age using the TW2 method. The identification and analysis of carpal bone regions of interest (CROI) is much more complex because of the intrinsic properties of the hand structure (high variability in density and morphology, contours not always visible etc.)*

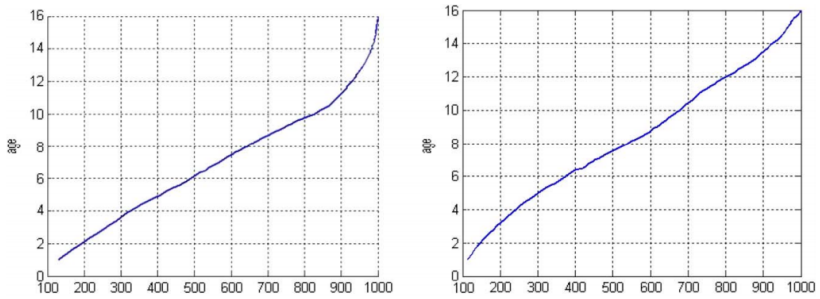

Figure 4.3: *Correspondence between TW2 final score and calculated age for males (Left) and females (Right).*

Preprocessing and Orientation Correction

Raw X-Ray images often contain some impurities and noise that make automatic processing difficult. For this reason it is necessary to enhance the clearness of the images and remove any unnecessary parts, such as background and radiological markers. As background we define an area outside the field of the radiation, which should be entirely black. In order to obtain a high hand-to-background ratio we employ a flood-fill approach [74]. After background removal, the next step is orientation correction. If the input image contains misaligned hands, it will be impossible, firstly, to identify the starting points for the algorithm, and, ultimately have an accurate assessment of the skeletal bone age.

To achieve this, we use wedge functions (i.e. binary functions aiming at identifying which part of an image belongs to the background and which one to the foreground [79], Fig. 4.7, left) to detect the middle finger and identify the axis (r_2) passing from the midpoint of the width at half-way the finger's length, and the midpoint of the width at its base (Fig. 4.5, Left). After identifying the angle of the r_2 axis with the vertical one, the image is rotated, obtaining the one shown in Fig. 4.5, right. The GUI of the tool showing the above operations is shown in Fig. 4.6

Finger Identification and EMROI Extraction

The next step after preprocessing of the input image, is to identify the fingers and extract the EMROIs from them. The TW2 method uses information contained in the first (thumb), third (middle) and

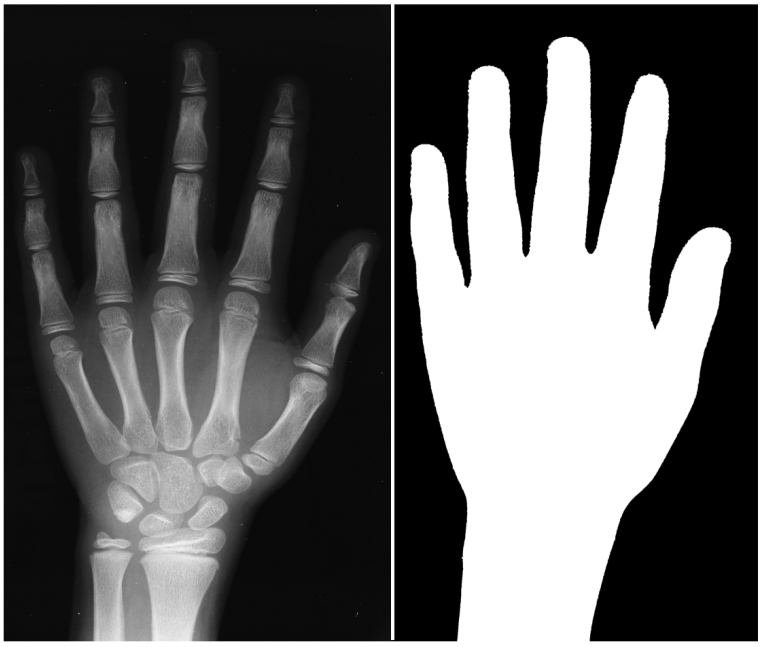

Figure 4.4: *Left: Original image. Right: The same image after the background removal algorithm was applied.*

fifth (pinkie) fingers. Identifying the third and fifth fingers is trivial. Once again, wedge functions are used in order to isolate each finger and find the tips, middle and base points of each one which correspond to the wedge functions' peaks. Each finger region is then rotated and extracted. Thumb extraction, instead, is implemented by a different procedure, consisting of identifying the points T_{thumb}, A and B (Fig. 4.7), by the following method:

1. We define as T_{thumb} the right-most, firstly, and top-most, secondly, point that belongs to the hand.

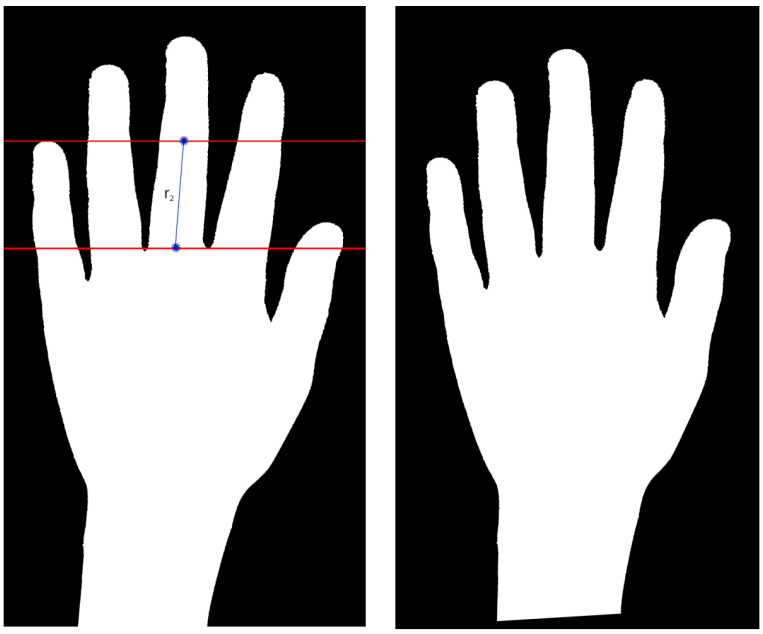

Figure 4.5: *Left: Identification of the r_2 axis. Right: The same image after orientation correction algorithm is applied.*

2. Starting from the T_{thumb} point, we scan from the X-Ray's right margin considering the wedges found line by line. When a non ending wedge is detected (i.e. the base of the thumb was found), we set as A the end point of the previous wedge found.

3. Starting from an arbitrary distance from the point A and scanning along the thumb's opposite side, we calculate the distance from A of all the points found and set as B the point that has the minimum one (Fig. 4.7, right).

The image is cut at the segment defined by the points A and B, which separates the thumb from the rest of the hand.

Figure 4.6: *X-Ray preprocessing: Through this GUI, the hand is preprocessed in order to remove unwanted signals and rotated in order to align it for further analysis.*

The EMROI extraction has been carried out by taking into account, for each finger, the average gray-level value as a representative for each horizontal line. Along this, we compute the first-order derivative to search for local maxima in the gray-level profile: these values indicate the bone borders for metaphysis, epiphysis, and diaphysis and are used to extract the EMROIs. We then apply, for reducing noise, smoothing filters on gray-level finger images. The first derivative is applied to the smoothed signal in order to enhance the EMROI.

Finally, by thresholding the previous signal, we extract the desired EMROIs. In fact, based on the peaks of the obtained filter, the distance between the middle and the distal part of the finger and the one between the proximal and the middle part of the finger are calculated.

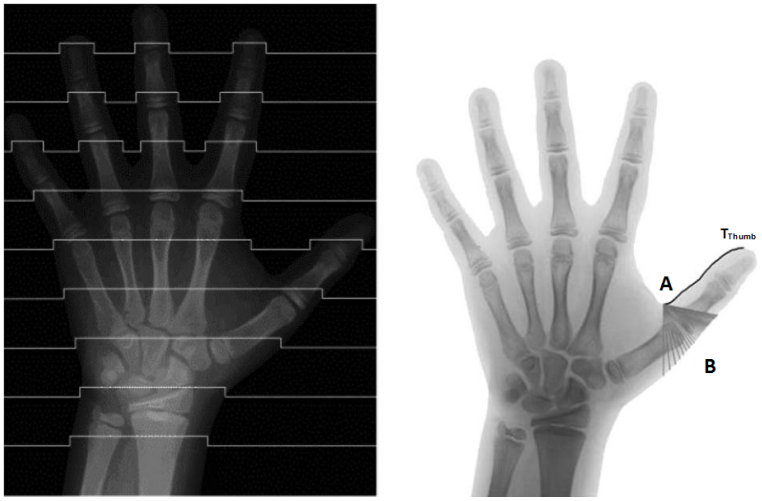

Figure 4.7: *Left: The wedge functions. The value of each wedge is equal to one if the underlying pixel belongs to the hand and zero otherwise. Right: Extraction and identification of the thumb*

If they are out of an anatomically plausible range, a warning message is displayed, and the procedure starts again by working on the derivative of the gray-level profile. The part of the application responsible for EMROI extraction is shown in Fig. 4.8

EMROI Enhancement and Feature Extraction

Once the EMROIs have been extracted, the *Difference of Gaussians (DoG)* filter [80] is applied for image enhancement. The DoG filter allows us to identify the soft tissue, typically appearing as a smooth region, by using a Gaussian function with a suitable standard deviation and, then, to remove it by subtracting it to another one with a less smoothing effect.

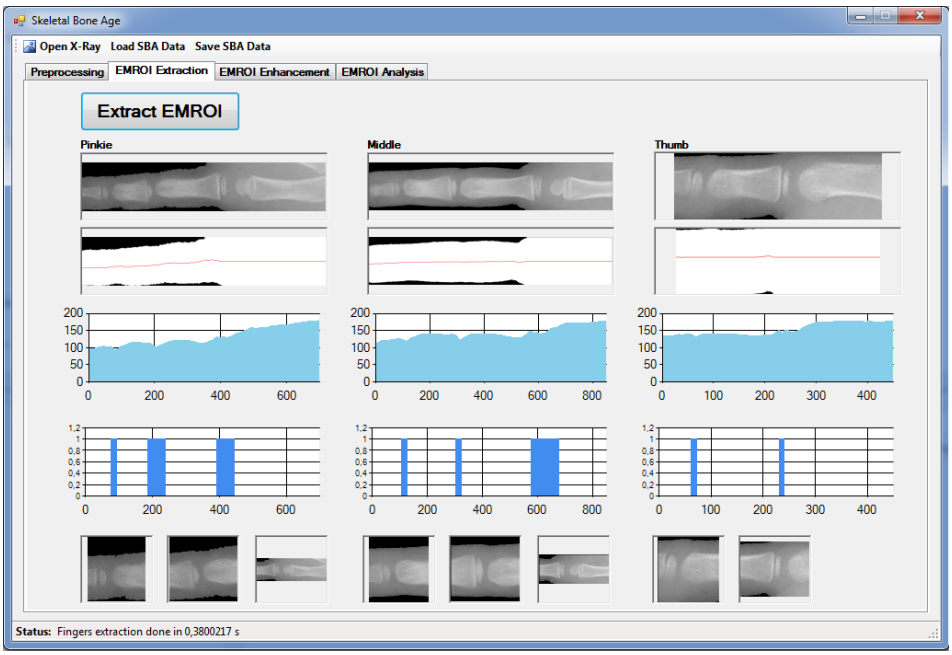

Figure 4.8: *EMROI Extraction: Following finger identification, in this GUI the single EMROIs are shown.*

By applying the DoG filter we remove all soft tissue, thus highlighting bone contours which permits their identification (Fig. 4.9, top right). The DoG filter, may leave bones shapes with holes and general shape degradation which can make impossible the extraction of the EMROIs and flood fill is applied if necessary(Fig. 4.9, bottom left).
In order to verify that the identified contours represent bone shapes we check for a set of geometrical and morphological criteria: metaphysis and diaphysis must touch the top and bottom margins of the image, respectively, and the epiphysis compactness must have a much higher value than the other ones. Additionally the relative values of

the areas of the identified bones should be in a fixed range of values and the finger's middle axis should pass through each bone.

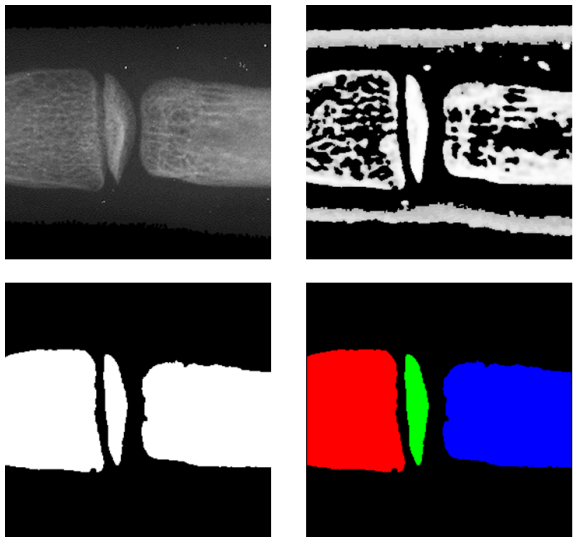

Figure 4.9: *EMROI Enhancement. From top left to right: 1) Original image 2) Image after the DoG fitler is applied 3) Enhanced image after filling the gaps 4) Identified EMROIs (red = metaphysis, green = epiphysis, blue = diaphysis)*

In case metaphysis and epiphysis are fused (i.e. only two bones were extracted, instead of three), the depth of the convexity defect is calculated in order to distinguish fused bones. The fused bones then are cut at the deepest points of such defect, making sure that no erroneous oblique cuts will occur.

When the number of extracted bones is equal to three the labels are assigned from left to right: metaphysis, epiphysis and diaphysis

(Fig. 4.9, bottom right). The procedure is implemented through the GUI shown in Fig. 4.11.

At this point, a thresholded image for each EMROI is available that contains the extracted bone. What remains is their identification, a task not always trivial because bad quality images may introduce undesirable effects in the form of fused bones.

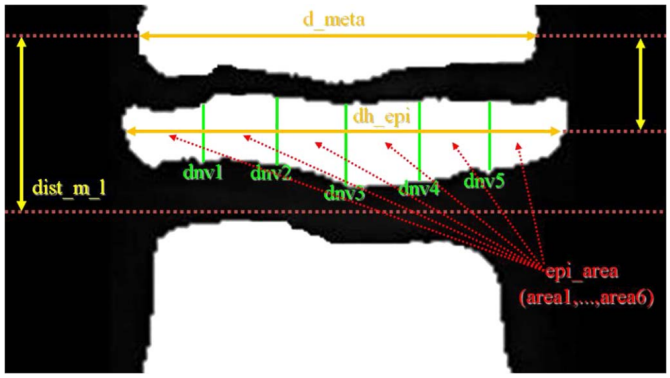

Figure 4.10: *Feature vectors for each EMROI*

The significant features of each detected EMROI for stage assignment purposes, are derived by the Tanner and Whitehouse (TW2) method. In detail, for each EMROI a feature vector containing the following features (shown in Fig. 4.10) is created:

$$FV_{bone_{stage}} = [d_{meta}, d_{m_1}, d_{nv_1}, ..., d_{nv_5}, d_{h_{epi}}, area_1, ..., area_6]$$

where d_{meta} is the width of the metaphysis, $d_{nv_1}, ..., d_{nv_5}$ are the heights of the different lines that divide the epiphysis' main axis in six equal parts, and $area_1, ..., area_6$ are their areas. Finally, $d_{h_{epi}}$ is the distance between the metaphysis and the diaphysis.

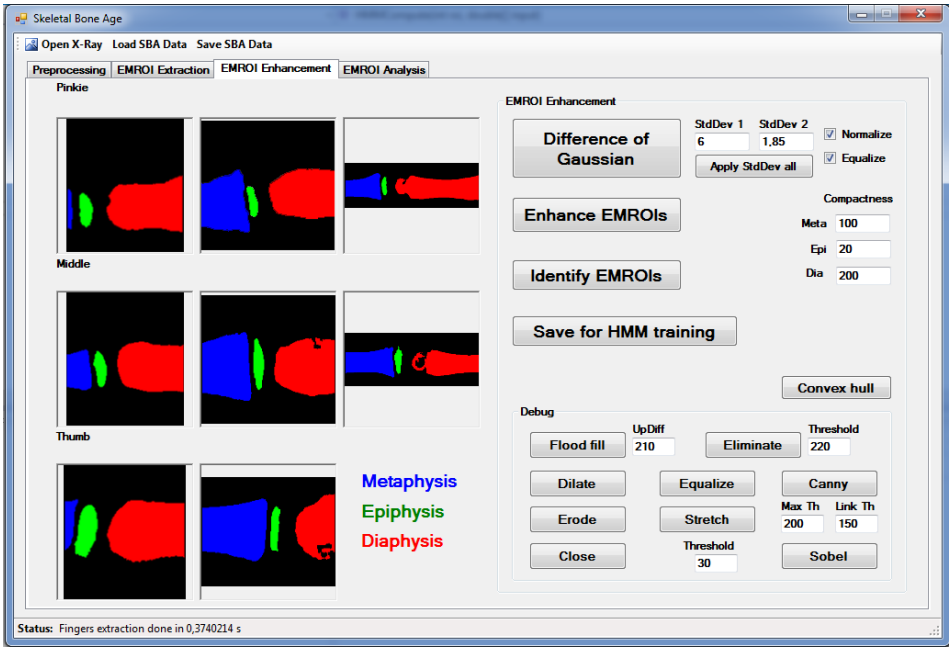

Figure 4.11: *EMROI Enhancement: In this GUI, the identified EMROIs are enhanced by applying the filters.*

Classification with Hidden Markov Models

The last module performs stage assignment and this is carried out by Hidden Markov Model (HMM) [81], which is a statistical model, similar to the regular Markov model with the difference that it contains only unobservable states. This means that the internal state of the model at any given time is not directly visible, but the output is. Each state of the HMM has a probability distribution over its outputs.

In our case we created an HMM for each EMROI and for each state (A-I), for a total of 72 HMM classifiers (27 HMMs for the middle finger, 27 for the pinkie and 18 for the thumb). The observations

of the models are the feature vectors described in the previous section. Given a feature vector of 14 elements it represents a set of 14 observations, one for each feature, that are introduced sequentially into the respective HMM, which returns the probability that the sequence belongs to it. In other words, the output of the single HMM are the probability that the fed feature vector was produced by it. Once these probabilities are calculated the HMM representing the state with the maximum likelihood is chosen and its respective stage label is assigned to the EMROI. The architecture of the HMM based classifier used can be seen in Fig. 4.12.

The output of this system is a set of letters representing the most probable developmental state for each EMROI. By summing up their corresponding values, the bone age of the subject is assessed (Fig. 4.13).

4.4 Experimental Results

The proposed method was tested with k-fold cross-validation method ($k = 5$). For the testing purposes we used 360 left-hand X-Ray images (180 males and 180 females), 30 X-Rays for each year in the range of $0 - 6$.

For each of the X-Rays the 14-value feature vectors were calculated and fed to the HMM models, and for each patient the developmental status of each EMROI was evaluated and then compared to the ones assessed by two expert physicians. While the importance of each single EMROI is important for assessing the accuracy of the system, calculating the discrepancy of the computed bone age with the effec-

tive one, in order to give a global idea about the performance of the whole system, is equally crucial.

In order to test how our approach performs compared to other approaches, we also tested the systems described in [74], in [82] and in [79]. The values for the $StdDev_1$ and $StdDev_2$ parameters of the DoG filter were set equal to 13 and 1.85, respectively, and these values were assessed empirically. In order to compare the performance in accuracy of all the systems we compared the results of all of them against the evaluations of the experts. A perfect evaluation is defined as the result where the developmental stage assignments correspond completely with the ones made by the experts. A good evaluation is defined as the result obtained when the maximum discrepancy from the expert's evaluation is of one stage and, finally, a bad EMROI assignment is considered when this discrepancy is of two stages or more.

As shown in Table 4.1, the proposed method outperformed all the others in terms of accuracy. In fact, assuming as correct evaluations the Perfect and the Good ones (i.e. the difference in EMROI stage assignment with respect to the radiologist's one does not exceed one step), the correct assignment rate is over 96% in both cases. The rest of the systems performed well (and in particular our previous work, [74]) but they did not, in any case, reach the performance of the proposed method.

While Table 4.1 reflects the performance of the systems in classifying correctly the single EMROIs, it should be noted that they do not express their potential in assessing the correct age of the individual. This aspect of the systems is shown in the Table 4.2 where the Mean Average Error (MAE) and the Standard Deviation (STD) between

Method	Radiologist	% Perfect	% Good	% Bad
Proposed Method	1	87.4	9.3	3.3
	2	91.1	8.2	0.7
Giordano [74]	1	82.8	8.5	8.7
	2	80.5	10.1	9.4
Pietka1 [79]	1	67.4	17.6	15
	2	68.1	13.8	18.1
Pietka2 [82]	1	71.2	17.1	11.7
	2	66.1	15.7	18.2

Table 4.1: *Number of EMROIs classified as Perfect, Good or Bad by all the systems with respect to the evaluations of the two radiologists.*

Method	Radiologist	MAE	STD
Proposed Method	1	0.37	0.29
	2	0.41	0.33
Giordano [74]	1	0.88	0.14
	2	0.61	0.22
Pietka1 [79]	1	2.63	0.93
	2	2.18	1.44
Pietka2 [82]	1	1.88	0.74
	2	1.98	1.07

Table 4.2: *MAE and STD values, showing the discrepancy (in years) between the tested methods and the radiologists evaluations.*

the computed bone age and the evaluations of the two radiologists are shown. Our method, achieved excellent results when compared both with the radiologistsévaluations (MAEs of 0.37 and 0.41, with STDs of 0.29 and 0.33, respectively). Such performance is clearly superior with respect to the other methods tested (best case MAEs of 0.88 and 0.61, with STDs of 0.14 and 0.22, respectively).

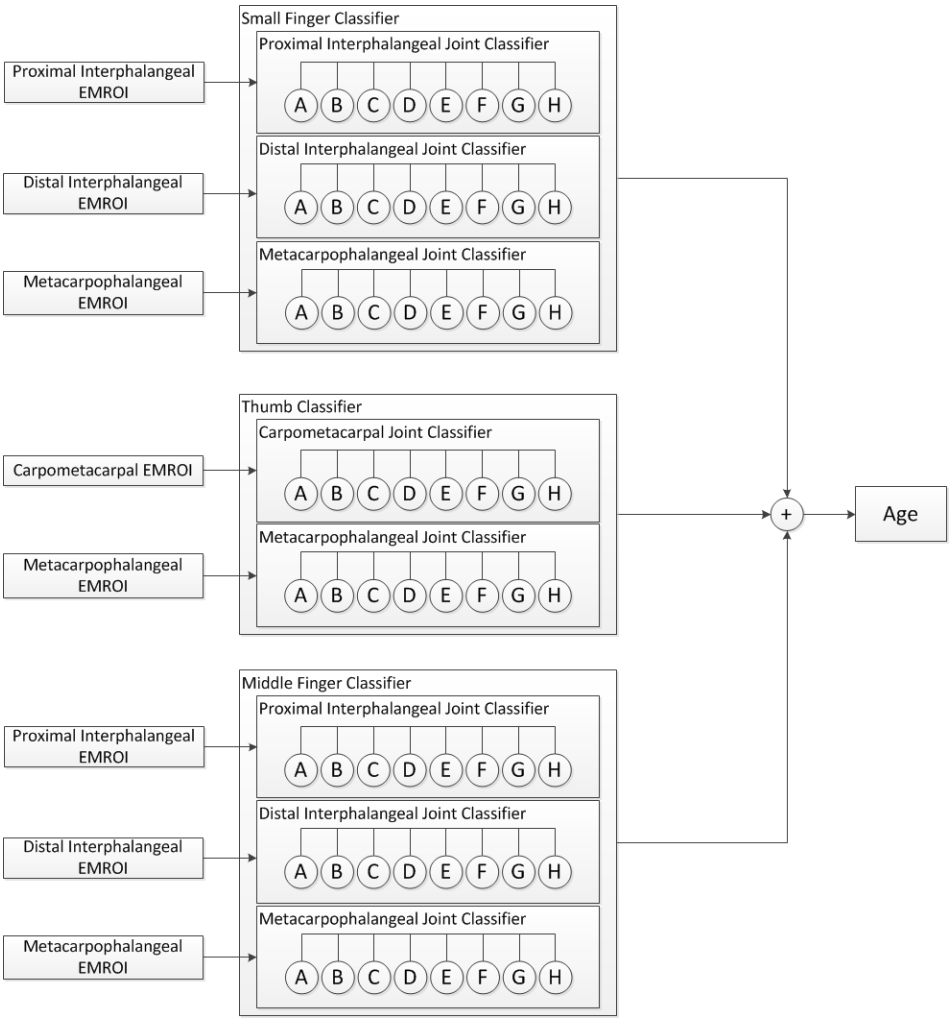

Figure 4.12: *The HMM architecture adopted in our approach. Each circle with a letter is an HMM that represents a developmental state. The EMROI classifiers have the role to select the maximum output of the underlying HMMs.*

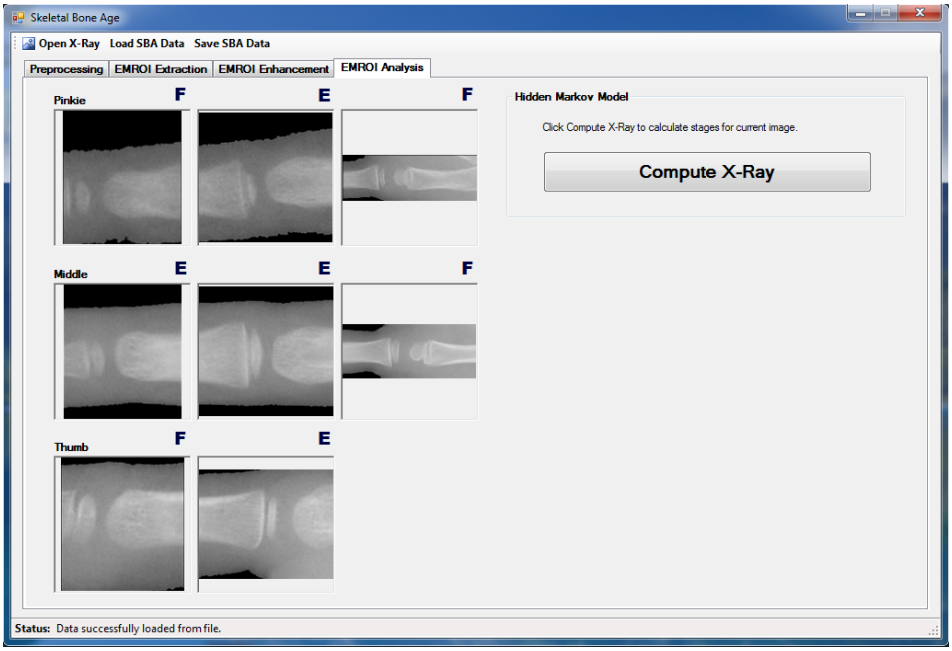

Figure 4.13: *EMROI Classification: When the processing terminates the HMM-based classifiers assign the letter of the developmental stage to each EMROI.*

CHAPTER

FIVE

KNOWLEDGE DISCOVERY IN THE MEDICAL DOMAIN

In this chapter, we present a bioinformatics knowledge discovery tool, BioCloud, for extracting and validating implicit associations between biological entities. The aim of this work is to demonstrate how porting a data-intensive application to the Cloud, affects positively its efficiency necessitating minimal effort. By mining specialized scientific literature, the proposed tool not only generates biological hypotheses in the form of associations between genes, proteins and diseases, but also validates the plausibility of such associations against high-throughput biological data (microarrays) and annotated databases. Both the knowledge discovery and its validation are carried out by exploiting the advantages and the potentialities of the Cloud, which allowed us to derive and check the validity of thousands of biological associations in a reasonable amount of time.

The proposed system employs a natural language processing (NLP) based approach for mining the scientific articles (not only abstract as in many state-of-the-art approaches) combined with existing standard vocabularies (MeSH, Entrez, UniProt) to infer associations, which are then filtered out according to their evidence in experimental data. Furthermore, starting from the valid associations, new associations are identified by mining only the experimental data.
The results shown that deploying the proposed tool in an IaaS cloud environment, a speed up of about 25% with respect to a locally running instance was achieved.

5.1 Introduction

A huge amount of biomedical information is hidden in millions of scientific articles published in the last 25 years and this quantity is exponentially increasing. This overwhelming quantity of information in the scientific literature compels, therefore, the need for new methodologies to discover new, previously unknown information available in the published papers in order to support biologists in their strive towards understanding/analysing biological data. One of the most effective and explored approaches to uncover this hidden knowledge is by mining the scientific literature [83, 84], especially for finding gene-gene [85], gene-disease [83] and protein-protein [86] associations. However, usually, the number of inferred associations (especially in the approaches which retrieve also first-order associations) can be massive, thus making the analysis and interpretation of such information as complex (and probably more cryptic) as reading all the scientific papers the

associations were extracted from. Therefore, issues such as validity, plausibility and feasibility of the inferred associations arise and, for this reason, methods, e.g. [87], to filter the obtained associations in order to distill (i.e. identify the most significant ones) the information and to propose it as significant scientific hypotheses have been investigated. A significant support to meet this goal comes from the massive publication on the Web of annotated chemical, genomic, clinical and other types of databases which could provide evidence and validate specific hypotheses. If, from one hand, experimental data may support the literature mining process, from the other hand, scientific literature may support the interpretation of such data, which often are extremely cryptic (e.g. list of up and down regulated genes)
.
Recent knowledge discovery systems, such as PathExpress [88], GenCLIP [89] and CoPub [90], ENDEAVOUR [91], GeneWizard [87], G2D [92], have exploited this integration between the literature and experimental data (biological, chemical, medical and drugs databases) for hypothesis generation and validation. Most of these tools integrate text mining and microarray data for extracting gene-gene and gene-disease associations and for gene prioritization, though approaches dealing with proteomics data [93] have been also proposed.

Therefore, digging out the "treasure" from massive biological data represents the primary challenge in bioinformatics with a consequent unprecedented demands on big data storage and analysis. In fact, with the amount of data growing and the increasing complexity of bioinformatics algorithms and tools, it is becoming highly demanding the introduction of advanced computational techniques to enable efficient knowledge discovery from data. However, it is often impos-

sible for small laboratories or even large institutions to establish and maintain large computational infrastructures for data processing. A promising and recent solution to address this need is Cloud Computing [94], which exploits the full potential of multiple computers and delivers computation and storage as dynamically allocated virtual resources via the Internet, thus representing an important alternative to ensure high performance data processing and easy management of complex tools in different areas of bioinformatics [95], data and text mining [96]. As a consequence of this, the number of cloud resources is increasing at an accelerating pace, with service-based cloud environments provided by Microsoft[1], Google[2], Amazon[3], SGI[4], and more, lending an unprecedented opportunity to evaluate the capabilities of the Cloud for sustainable and large-scale data processing in bioinformatics.

Typical uses of the Cloud are mainly in the areas of economics, health, and the entertainment industry, whereas its application in bioinformatics has been mainly oriented to the field of comparative genomics, e.g. the Sanger Institutes fast matching and alignment algorithm to assemble full human genome [97], Cloud Burst [98] to map next generation sequencing data [99], Cloud Blast a "clouded" implementation of NCBI BLAST [100, 101]. However, other bioinformatics approaches exploiting the potentialities of the Cloud have been proposed recently and will be reviewed in the next section. Therefore, bioinformatics is experiencing a new leap-forward: from in-house computing infras-

[1]http://www.windowsazure.com/en-us/
[2]https://cloud.google.com/
[3]http://aws.amazon.com/ec2/
[4]http://www.sgi.com/solutions/internet/

tructure into utility-supplied cloud computing delivered over the Internet, in order to extract knowledge from the vast quantities of biological data generated by high-throughput experimental technologies and from the huge bulk of available scientific papers. In addition, the Cloud does not offer only computational infrastructure for large scale data processing, but also a set of services which can be exploited for speeding up the research on bioinformatics.
The main contributions of this work to the research on bioinformatics are:

- A systematic review of the existing cloud based services, approaches and tools in bioinformatics;

- One of the first examples of cloud based knowledge discovery system, BioCloud, which generates biological hypotheses in the form of gene-disease relationships by mining scientific papers and then validates the inferred associations against microarray data.

The remainder of the chapter is as follows: Section 2 reviews the existing Cloud services and infrastructure that might be adopted in the bioinformatics research; Section 3, instead, describes BioCloud, a knowledge discovery tool that employs a NLP based for mining the literature and deriving associations between biological entities, which are further validated against high-throughput data. Since the processes of text and data mining are expensive in terms of computational resources and processing times, BioCloud uses Cloud Foundry, a platform for development, deployment, and operation of cloud applications. In Section 4 some experimental results of BioCloud are given.

5.2 Cloud Technologies in Bioinformatics

The rise of Cloud technologies represents an important and incredible opportunity for bioinformatics in order to satisfy the needs of processing large amounts of heterogeneous data, of storing massive amount of data and of using the existing tools in different fields of bioinformatics. This importance is witnessed by the ever growing number of bioinformatics applications (from DNA sequencing [102], to sequence alignment and similarity search [103], data mining [104] and knowledge discovery [96]) relying on Cloud services. However, Cloud computing does not serve only for large scale computation but it is changing radically the traditional way of doing research leading to a new era of bioinformatics. In fact, the typical workflow foresees that biologists design the experiments and send samples to sequencing centers, which make available raw data (through specific services, such as FTP, HTTP) to biologists, who have to download in their local institutions terabytes of data and, according to the research plans, publish this data in public databases. At the same, biologists copy the data into local machines for being used by bioinformaticians for the subsequent data analysis. Bioinformaticians, on the other hand, when process biologists' data have also to download data from public databases. Therefore, this typical flowchart (see Fig. 5.1) implies that huge quantities of data are moved several times from sites to sites, thus slowing down the analysis and the interpretation of the results.

The cloud, instead, aims at creating an infrastructure (see Fig. 5.2) where sequencing centers store their data into the cloud, public databases are built on the top of the infrastructure, biologists access this data directly from the cloud and share what they need with bioin-

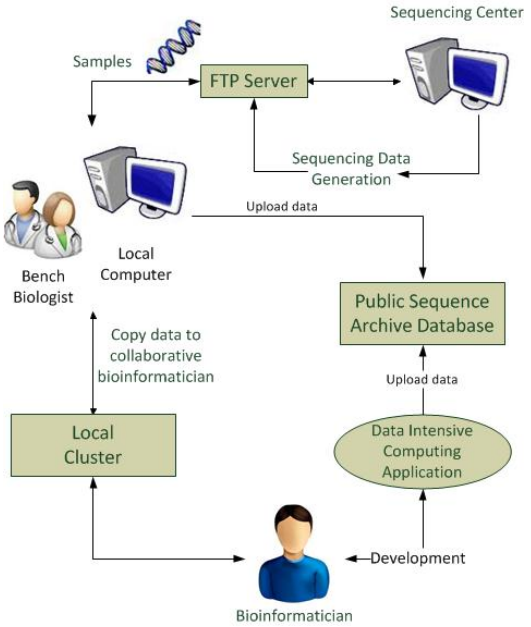

Figure 5.1: *Example of the typical workflow in bioinformatics.*

formaticians, who will develop large scale applications directly on the cloud whose results will be made available to the biologists for the interpretation. This new architecture will reduce the times data are transferred, but also it will allow laboratories and institutions to cut down the expenses to carry out experiments and data analysis.

In next the sections the existing Cloud services and solutions will be reviewed according to its service model categorization: Platform as a Service (PaaS), Software as a Service (SaaS) and Infrastructure as a Service (IaaS).

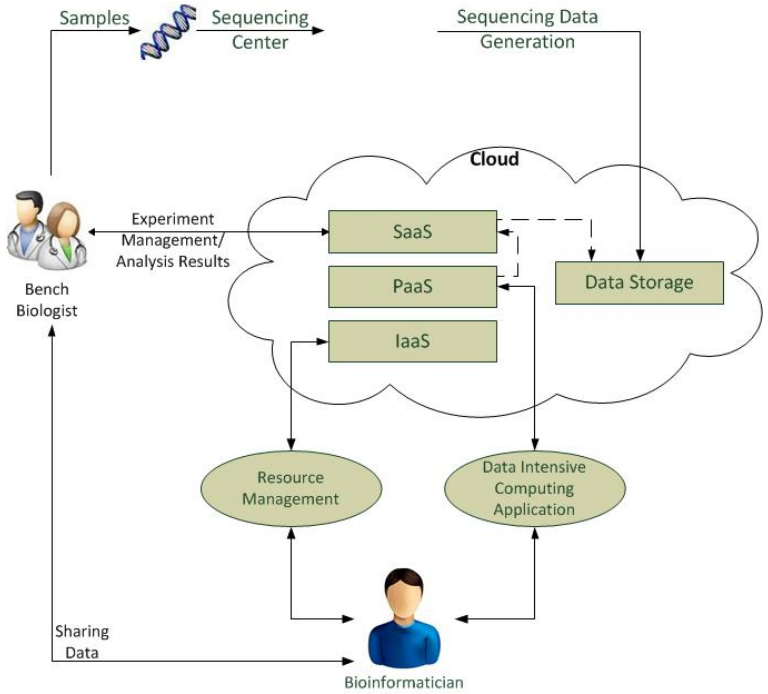

Figure 5.2: *Changes on the bioinformatics workflow with the introduction of the cloud computing*

5.2.1 Platform as a Service (PaaS)

Platform as a Service offers a development environment that allows users to create and run their applications using specific programming languages and frameworks available in the platform itself. Examples of PaaS environments are Google App-Engine[5] and Microsoft Azure[6]. However, to perform large-scale data analysis in bioinformatics it is necessary that Cloud based environments support the communication

[5]http://code.google.com/appengine/
[6]http://www.microsoft.com/windowsazure/

of parallel tasks in order to make full use of the available computation and storage resources. To address this need, most of the existing PaaS services are provided with an additional abstraction level implementing the map-reduce programming model. The map-reduce computational paradigm divides the main application into many sub-applications, each executed or re-executed on a node of the Cloud infrastructure, and consists of two main steps. During the first step (map), the master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. The worker nodes process the smaller problems, and pass the answer back to its master node. In the second step (reduce), the master node collects the answers to all the sub-problems and combines them to form the output. There exist many frameworks implementing the map-reduce paradigm that also provide jobs management functions for data-intensive computing such as Apache Hadoop[7] or Microsoft's Dryad[8]. More in detail:

- Apache Hadoop framework, which beyond the implementation of the map-reduce model, provides a distributed file system, the Hadoop Distributed File System (HDFS) [105], for effective and very low latency data storage on the worker nodes. In addition, there are many projects built on top of Hadoop such as Pig[9] which is a high-level data-flow language and execution framework whose compiler produces sequences of Map/Reduce programs for execution within Hadoop, or Hive [106] which is a data warehouse framework built on top of Hadoop, developed

[7]http://hadoop.apache.org/
[8]http://research.microsoft.com/en-us/projects/dryad/
[9]http://pig.apache.org/

at Facebook, used for ad hoc querying with an SQL type query language and also used for more complex analysis.

- Microsoft Dryad, developed by Microsoft Research, allows developers to write parallel applications executing on the Cloud by modeling a directed acyclic graph (DAG). The DAG consists of a set of vertices describing the operations to be performed, which are distributed at runtime to different execution engines.

- Cloud MapReduce [107], is an implementation of MapReduce model [108] on top of the Amazon Cloud OS. Cloud MapReduce can be considered as an optimized version of the other MapReduce implementations, thanks to an architecture that ensures several advantages in terms of speed, scalability and simplicity.

Recently, PaaS frameworks have been applied with increasing interest to bioinformatics research as demonstrated by the quantity of works employing the map-reduce approach on the Cloud, mainly, for parallel large scale data processing. In [109], Windows Azure was used in particular for data storage and as VM (Virtual Machine) hosting environment to conduct data mining for computational drug discovery. In [110], Dryad and Hadoop were used to host two bioinformatics applications: Expressed Sequence Tag and Alu Sequencing. An accurate performance evaluation showed the advantages of the two frameworks with respect to traditional MPI implementations.
To the best of our knowledge there exist only few applications exploiting cloud based map-reduce solutions to perform literature text mining for biological hypothesis generation. Nazareno *et al.* in [111] proposes an "ad-hoc" cloud infrastructure for identifying molecular interactions

by mining the scientific literature, whereas Lin *et al.* in [112] describe, more generally, how to process massively text data by using MapReduce. However, these solutions are at a very early stage and their implementations can not be used reliably for massive text processing due mainly to the lack of generalization. In fact, the deployment of these systems is too application-specific and often restricted to single private cloud environment.

Unlike text processing, much more cloud based map-reduce methods (mainly based on Hadoop) have been proposed for processing high throughput data analysis. Crossbow[113] proposes solutions executing on Hadoop for whole genome resequencing analysis and SNP genotyping from short reads. Contrail [114] uses Hadoop for de novo assembly from short sequencing reads, whereas Myrna [115] proposes a method for calculating differential gene expression from large RNA-seq data sets. On clusters Myrna uses Hadoop, whereas in the Cloud it uses Amazon Elastic MapReduce[10]. Analogously, a few Cloud-based methods for microarray data mining analysis have been proposed as in [116] where the authors developed a MapReduce framework on Hadoop for mining association rules from microarray gene expression datasets. Delmerico *et al.* in [117] provide an extensive performance evaluation of clusters and Hadoop based solutions for computing genes correlations by processing microarray data. The authors state that although the performance of the existing approaches for identifying such correlations are generally improved on clusters, storage, hardware and network (mainly) limitations restrict their scalability, on the contrary of Hadoop, which, instead, provides a significantly better scalability.

[10]http://aws.amazon.com/elasticmapreduce/

However, two are the main downsides of MapReduce solutions: first of all, the map/reduce frameworks require re-writing most of the existing applications, which, for several reasons, is not appreciated by bioinformaticians and biologists. Second, the current implementations of map/reduce paradigm employ some overly simple mechanisms; for example, the job scheduling is often not (well) supported, thus affecting the tools' performance.

5.2.2 Software as a Service (SaaS)

Software as a Service is a solution which delivers the software applications online, thus facilitating remote access to available bioinformatics software tools through the Internet. In SaaS services there is no client side software requirement for the user: the services are reachable through an access point like a web portal or a visualization tool. The main advantage of SaaS is that it enables large scale data analysis over the web, thus eliminating the need for local installation of a large variety of software tools and also providing up-to-date cloud-based services for bioinformatic data analysis.

An interesting example of SaaS is EasyGenomics[11] a key enabling platform providing streamlined bioinformatics services. Basically, most of the available tools implement the map/reduce paradigm for parallelization and scalability, but make it trasparent for the end-user who have to call only the service without bothering about the underlying software and hardware infrastructure. Relevant examples are: CloudAligner [118] a full-featured Hadoop MapReduce-based tool for

[11] www.easygenomics.com

sequence mapping and CloudBurst [98] an open source optimized tool for mapping next-generation sequence data to the human genome with MapReduce.
However, one of the major requirement of SaaS services in bioinformatics is the interoperability between multiple cloud systems. The main difficulties to address this need are that the mechanisms for service publishing, searching and subscription are not well-established and the existing technologies (WSDL, UDDI) do not describe sufficiently the semantics of such services. For this reason, the current trend is towards a metadata ontology that would help to describe the service metadata including service providers.

5.2.3 Infrastracture as a Service (IaaS)

The infrastructure-as-a-service (IaaS) layer aims at offering computer infrastructures, virtualized resources, storage, networks, and other fundamental computing resources via self-services to the user. The challenge introduced by bioinformatics on IaaS regards the enhancement of flexibility of cloud platform for resource management in order to satify user needs. The most appropriate approch to ensure such flexibility is via virtualization that mainly involves either the generation of multiple virtual machine instances to partition the physical resources or multi-tenancy techniques, which enable users to share application instances and treat them as independent ones.
However, currently, the most employed approach is the creation of suitable virtual machine instances according to user requirements. This is a non-trival task in bioinformatics because of dependency and version matching issues arising when dealing with bioinformatics tools.

Amazon EC2 [119] represents the first example of such a service and it offers a variety of VM images provided with a good variety of bioinformatics tools. Other important examples are: Cloud BioLinux [120] and CloVR [121]. The former is a publicly accessible virtual machine for high performance bioinformatics computing. The latter, instead, is a portable virtual machine for automated sequence analysis and its performance are discussed in [122, 123].
The main limitation of the current IaaS services is that VM creation, update and sharing is too ad-hoc and taloired to the specific needs of bioinformaticians and biologists, who, basically, have to create VMs from the beginning. Recently, on-demand packaging mechanisms are under investigation to allow an automatic creation of virtual machine images provided with all the needed and up-to-date tools with all the dependencies solved.

In the next section, BioCloud, a knowledge discovery tool for identifying gene-disease associations exploiting an IaaS cloud-service is described as an example on how to execute large-scale data analysis tool on the Cloud.

5.3 BioCloud

BioCloud is an application that allows users to produce new biological hypotheses through an intuitive and guided interface without requiring knowledge of text-mining and data-mining methods. It retrieves automatically gene-disease (but also gene-gene, protein-protein and protein-disease) associations by mining Pubmed abstracts and Oxford Journals full papers and validates them against microarray data. Since

the associations retrieval and validation involves large scale text and data mining procedures, we have used Cloud Foundry[12] a cloud computing PaaS and IaaS solution developed by VMware[13].
In detail, the steps performed by BioCloud for hypothesis generation and validation are:

1. Text mining of scientific papers which involves three tasks:

    - Document Retrieval and dictionaries building. A dictionary of genes (Entrez Gene), a dictionary of diseases (MeSH) and a set of biomedical scientific abstracts and papers (PubMed and Oxford Journals) are used as basis of our text mining approach. Fig. 5.3 and 5.4 show respectively, the GUIs for document retrieval either from Pubmed or from Oxford Journal and for protein dictionary building.
    - A named entity recognition module which aims at improving the dictionaries' creation. In fact by only using the terms of standard vocabularies (e.g. Entrez Gene) it may happen that no association is derived because of the dissimilarities between the vocabularies' terms and the terms extracted from parsing full papers and abstracts;
    - A Natural Language Processing based approach that analyzes syntax and semantics of the retrieved papers for obtaining relationships between the entities of aforementioned dictionaries.

[12]http://www.cloudfoundry.com/
[13]http://www.vmware.com

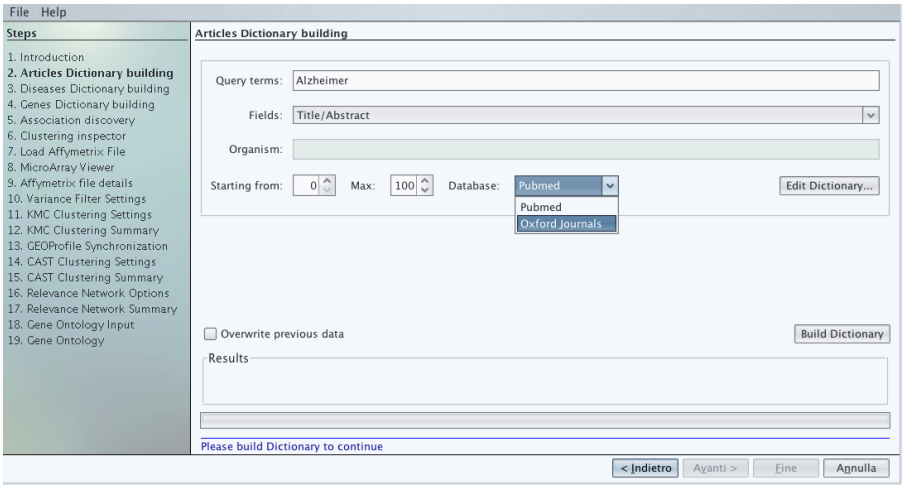

Figure 5.3: *BioCloud's GUI for document retrieval*

2. Validation of the derived associations using microarray data (gathered from the public GEO database, see Fig. 5.5) for gene-disease associations.

3. Execution of the algorithms on the Cloud. The modular architecture of BioCloud allows us a parallel execution on CloudFoundry of each module from document retrieval to natural language processing to microarray data analysis.

Fig. 5.6 recaps the resources and modules used by BioCloud. In the next subsections each module is described in detail.

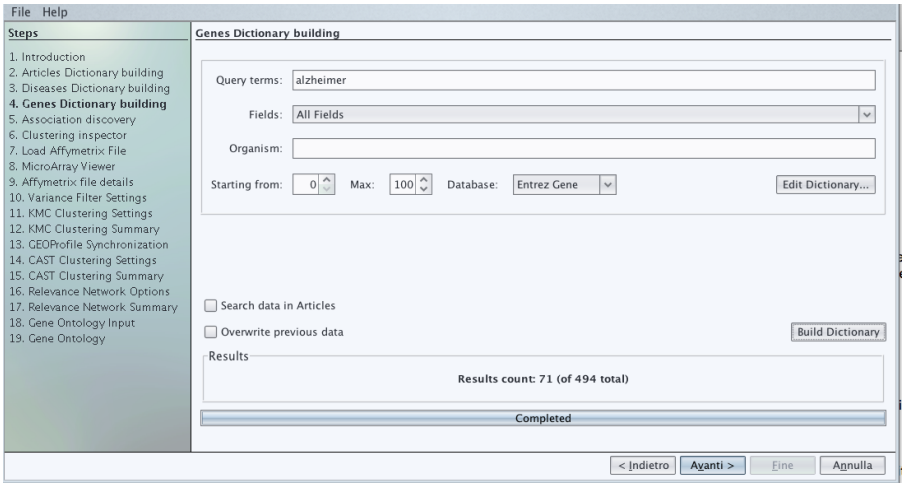

Figure 5.4: *BioCloud's GUI for gene dictionary building*

5.3.1 Text Mining Module for Hypothesis Generation

The text mining approach implemented in BioCloud is based on a natural language processing method which parses fully syntax and semantics of the retrieved papers. BioCloud infers an association between two biological entities $T_1 - T_2$ when it finds a meaningful triple (NOUN-VERB-ADJECTIVE) with NOUN and ADJECTIVE being genes/diseases names (taken from the biological terms vocabulary) and VERB being a verb which significantly correlates the two biological entities (e.g. T1 activates T2). In a previous work [87] we employed co-occurrences processing for deriving associations, that, unlike the one herein proposed, produces a lot of noisy associations (high recall, but low precision) making the subsequent validation very time comsuming and sometime also useless. The proposed approach

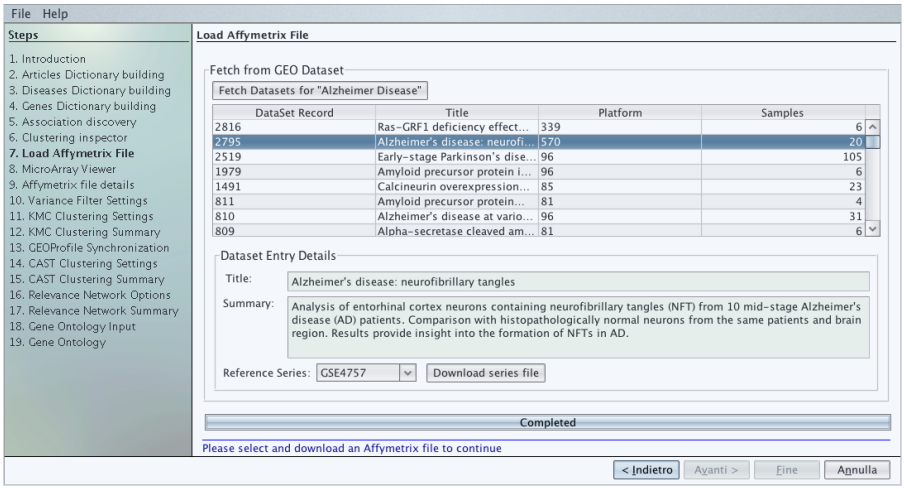

Figure 5.5: *Microarray data retrieval using Alzheimer disease as query term. This data is then clustered and a Gene Relevance Network is achieved.*

consists of four main steps:

1. **Document Retrieval and Dictionaries Building**. Since biological entities identified by mining only abstracts are underestimated because of abstracts' concise nature, the proposed approach uses a set of full text articles retrieved from the Oxford Journal system using a disease name as query term. At the same time, since we are dealing with gene-disease associations a dictionary of genes from Entrez Gene and a dictionary of diseases from Mesh are created.
   The creation of the disease's dictionary is restricted to the C branch of MeSH that contains only the classification of diseases.

2. **Natural Language Processing for Parsing Full Text**. In

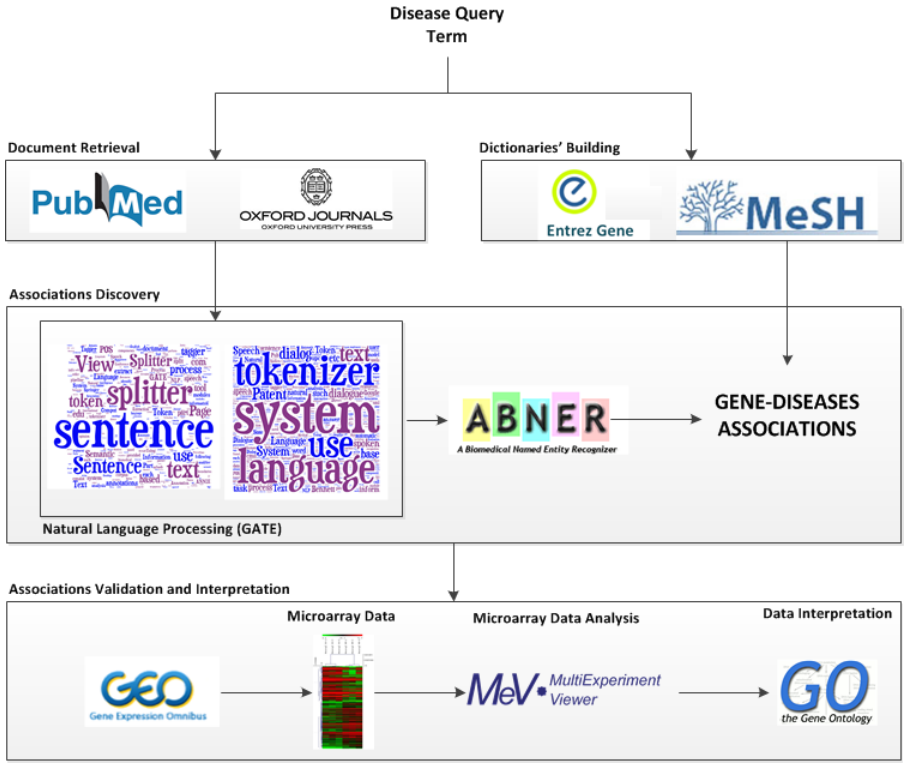

Figure 5.6: *Outline of the modules and resources used in BioCloud*

parallel to dictionaries building, the retrieved papers are parsed by using the ANNIE (a Nearly New Information Extraction System) module included in GATE [124]. The text parsing consists of the following modules: 1) *Text Tokenizing* to break the text into tokens, which provide useful information such as token category (proper noun, verb,adjective), token length and orthography (hyphenation, capitalization, word breaks) and 2) *Sentence Splitter* to split tokens into sentences.

3. **Named Entity Recognition**. We used ABNER [125] to tag the biological entities within a sentence.

4. **Associations discovery between meaningful terms**. For each sentence all the triples (NOUN, VERB, ADJECTIVE) are detected, and then, only those containing valid biological entity names (the ones in the dictionaries and validated by the NER) and consistent verbs (previously defined) are considered as hypotheses which are subsequently validated with microarray data.

BioCloud, moreover, allows the users to re-use the inferred associations in different mining processes in order to achieve multiple first order associations: i.e. if in a mining process we obtain an association between the gene G_1 and the disease D_1 and in another mining process we infer an association between the gene G_1 and the disease D_2, then a graph is created with a connection between D_1 and D_2 through the gene G_1. Fig. 5.7 shows the case of multiple associations between diseases and genes, whereas Fig. 5.9 shows multiple associations between proteins.

5.3.2 Validation of Hypothesis Generation against Experimental Data

For the validation of the generated hypotheses (gene-disease associations) we use microarray data from the GEO database using a disease-gene pair of a specific association. Once a microarray is selected, the tool starts data analysis in order to construct the relative gene relevance network (GRN) (i.e. a list of relevant genes for the

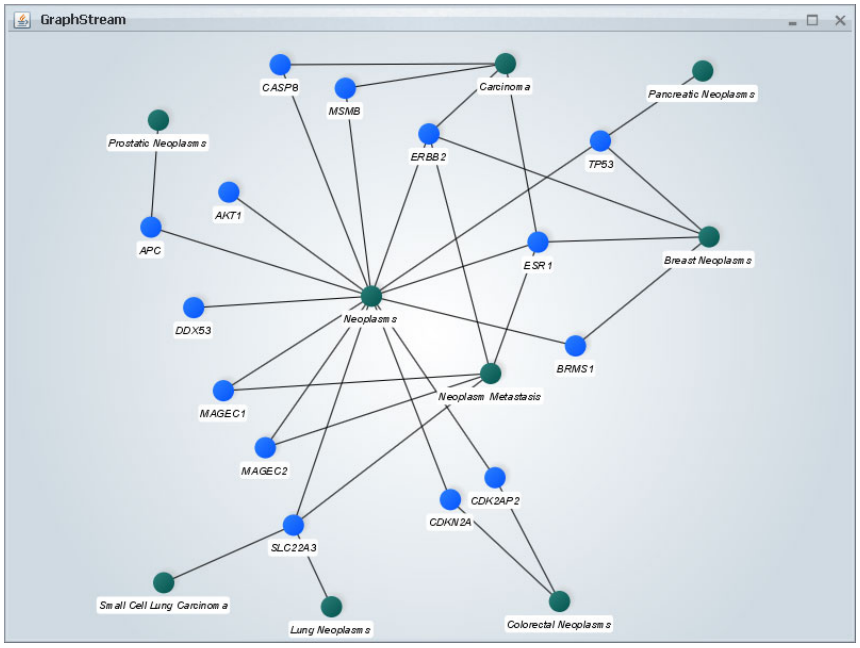

Figure 5.7: *Multiple disease-gene associations. Diseases are green-coded, while genes are blue-coded.*

given disease) containing the gene of the gene-disease association to be validated. The genes of the GRN are then re-codified using DAVID [126]. The microarray analysis modules are based on the Java classes from the MEV (MultiExperiment Viewer) software [127]. The first step of the data analysis is to apply Hierarchical Clustering to the microarray data to obtain clusters of genes. Then the cluster containing the gene under examination is selected and a GRN ("Main GNR") is derived from it by applying Cluster Affinity Search [128]. Since, often, the GNR may not contain a sufficient number of genes due to several factors (ranging from the clustering settings to the

microarray data quality) we iterate the procedure on building GNR for all the genes that make part of the "Main GNR". In detail, for each gene of the GRN, the microarray datasets that contain it are downloaded from the GEO database and then, according to the described procedure, another gene relevance network that will be connected to the main GRN (i.e. the one containing the disease under investigation), is built. Finally, the genes of this extended GRN are used to query the Gene Ontology (GO) database in order to investigate the biological meaning of such genes with respect to the given disease.

It is understandable that this validation process may not be executed on a single machine since it involves, first, a computational intensive text mining procedure and, second, a recursive data mining phase (several clustering steps executed on matrices with thousands of elements) for building the extended GRN.

Beyond gene-disease associations, BioCloud allows users also to extract protein-protein and protein-disease associations as shown in Fig.5.8. The main differences are the use of 1) UniProt[14] to build the proteins' dictionary and 2) UniProtJAPI[15] for implementing the NER. An example of derived protein-protein associations is depicted in Fig. 5.9.

The validation of protein-protein associations is, instead, performed by using BioContrasts[16] and STRING 9.0[17], which is a database of known and predicted protein interactions that makes use of

[14]http://www.uniprot.org

[15]http://www.ebi.ac.uk/uniprot/remotingAPI/

[16]http://biocontrasts.biopathway.org/

[17]http://string-db.org/

genomic and high-throughput experiments data. In detail, each identified protein-protein association is first passed for validation to BioContrasts which identifies contrasts between proteins by identifying patterns in the form of "A but not B" in MEDLINE abstracts. If there is not a contrast between the two proteins, the final validation step is to check if the association exists in STRING 9.0, which also provides a set of further proteins involved in the association. Similarly to the case of gene-disease association, we use the Gene Ontology to investigate the biological meaning of the proteins previously identified.
Protein-disease associations are instead automatically validated against the Human Protein Atlas[18] and the Human Protein Reference Database[19].

5.3.3 Data Analysis on the Cloud

CloudFoundry[20] offers the PaaS and IaaS cloud service models. The whole platform is controlled by a command line utility, called VMC, which allows the user to customize the hardware and software components that fit her needs. In CloudFoundry several database engines can be bound to each deployed application. In our case, each application instance was assigned two CPU cores, 2GB of memory and 1GB of hard disk space. For data management, a MySQL database combined with the Hibernate library was chosen. The cloud

[18]http://www.proteinatlas.org
[19]http://www.hprd.org
[20]http://www.cloudfoundry.com/

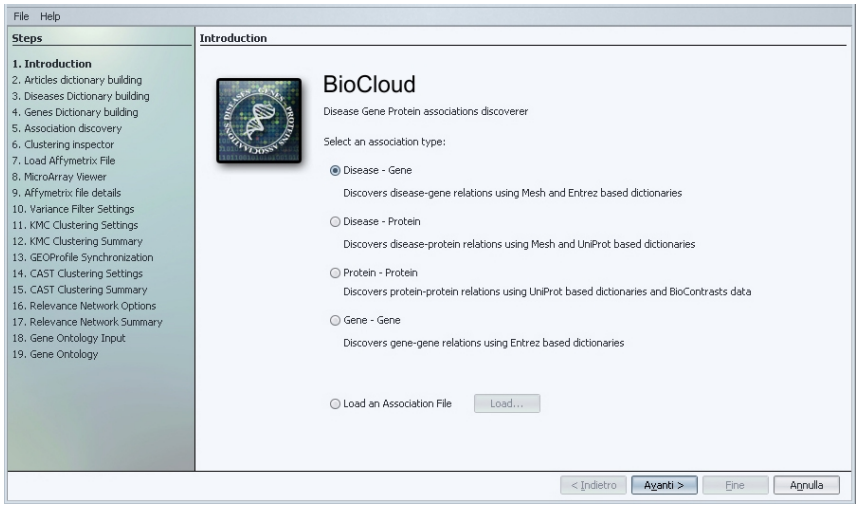

Figure 5.8: *Biological entities associations supported by BioCloud.*

execution requires to remove the interface part, thus we only execute the application engine giving the needed parameters/settings in a XML file.

As soon as the platform is set up, the application launches. The execution time is monitored by the VMC itself but the running application's standard output can be checked. When the program's execution ends, all the retrieved and produced data can be found in the database. This data can also be used in order to derive performance parameters (such as recall, efficiency metrics etc...).

Despite the fact that the virtual machine's hardware specifications were not top-performing, as shown from the experimental results, this configuration achieved more than adequate performance.

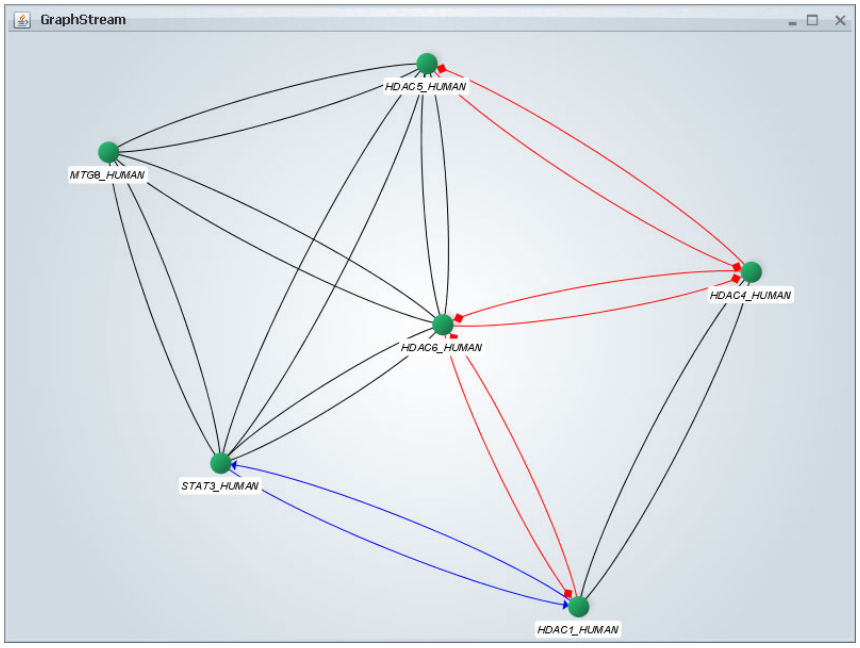

Figure 5.9: *Multiple protein-protein associations. The associations are divided into positive ones, and contrastive ones and are depicted with different colors in the graph.*

5.4 Experimental Results

As its name implies, knowledge discovery systems often produce results that are based on true scientific basis but not always hold true. This means that performance analysis of knowledge discovery systems is quite approximate given that it is not always easy to differentiate a valid discovered association from an invalid one. Therefore, the only reliable way to examine accurately the performance of a knowledge discovery system is to employ gold standard annotations and then

Number of diseases	110
Number of articles examined	220782
Number of relevant associations	1318

Table 5.1: *Synthesis of the dataset used for our experimental evaluation.*

compare the obtained associations against them.

The performance evaluation of BioCloud was divided in two main sections: 1) comparison, in terms of valid (i.e. that have evidence in experimental data) retrieved associations, between the NLP method, and the co-occurrence-based approach [87] for reference, 2) assessment of how the cloud implementation combined with the NLP method affects the efficiency of the operation. The local computer that we tested the tool on, had an Intel Core 2 Duo CPU running at a frequency of $2.67GHz$ with 4GB of RAM, while the cloud computing instance was the one described in Subsection 3.3.

As a gold standard (Table 5.1) we used the list of 110 diseases with 1318 associations to genes, described in [129]. For consistency, only the diseases with a minimum count of 100 retrieved documents were considered. For the totality of the diseases the application retrieved 220782 articles.

The time needed to complete the processing of this articles in each setting is shown in Table 5.2. It is clear how the cloud infrastructure used boosts the performance of the application. In fact, comparing it against locally running instance we observed a speed up in efficiency of about 25% when the NLP module was used (1715 against 2281 minutes) and about 26% when the co-occurrence module was used (3207

	Local		Cloud	
	NLP	Co-occurrence	NLP	Co-occurrence
Time needed in minutes	2281	4307	1715	3207

Table 5.2: *Performance in terms of time needed for each setting to process the whole stack of documents.*

minutes against 4307). From the same table we can also observe that the cloud-based solution offers a net speed up, when the NLP implementation was used instead of the co-occurrence method, of about 47%.

It is important to notice that the overall speed-up of the cloud-based implementation is due to the fact that file transfer and IO operations are done between entities that are on the cloud, meaning that a cloud-optimized NLP module would perform even better in this context.

Table 5.3 shows the achieved results of the cloud implementation in terms of valid associations/retrieved associations and recall. Under this aspect both the solutions performed identically because the algorithm's logic was identical. Observing the results it can be deduced that consistently better recall values are achieved when the NLP module is used. While the PA parameters (in Table 5.3) could be considered as False Positive values (i.e. associations that did not exist in the gold standard dataset but the tool found the opposite), it represents possible hidden information, that must be further investigated.

Disease	RA	DA	Co-occurence					NLP				
			TP	PA	RC	TL	TC	TP	PA	RC	TL	TC
Anemia	5810	33	20	18	0.61	10	7	26	21	0.79	4	3
Breast Cancer	10000	24	15	13	0.63	219	165	21	16	0.88	62	49
Diabetes Melitus	10000	38	13	19	0.34	190	143	28	28	0.74	123	91
Hypertension	10000	13	7	2	0.54	202	142	9	3	0.69	98	73
Leukemia	10000	39	20	16	0.51	233	178	24	26	0.62	146	113
Liver Cancer	8175	10	4	4	0.40	177	125	6	7	0.60	114	82
Lymphoma	10000	10	6	7	0.60	225	159	6	9	0.60	133	103
Melanoma	7931	6	3	7	0.50	181	143	4	4	0.67	113	84
Obesity	10000	24	15	15	0.63	145	114	17	21	0.71	110	86
Prostate Cancer	7652	14	5	8	0.36	120	85	11	8	0.79	76	58

Table 5.3: *Experimental results of a subset of the dataset in terms of valid associations/retrieved associations per disease, average precision and recall. RA is the number of the retrieved articles from Oxford Journal when the corresponding diseases name was queried, DA denotes the number of the existing associations in the gold standard dataset. TP is the number of gene-disease associations that were both in the gold standard dataset and the applications output and PA denotes the number of associations that did not exist in the gold standard dataset but the tool marked them as valid (Possible Associations). RC represent the recall for the corresponding disease. TL and TC is the time, in minutes, needed in order to complete the processing, respectively, for the local instance and the cloud.*

CHAPTER

SIX

CONCLUSIONS

Medical data come in many formats and each format need completely different thinking and processing. During my Ph.D course, I identified four distinct problems that the medical world needs to deal with and tried to tackle them using a multitude of methods.

In Chapter 2, a system that automatically generates summaries taking as input the corpus of unstructured medical reports, was presented. Such summaries, are also annotated with links which the reader can follow in order to get a short description of the corresponding medical concepts achieving very good performance. The same system could be configured to use the International Classification of Diseases (ICD) dictionary, instead of or in addition to UMLS, to assign codes to diseases making the system more compatible with existing systems.

In Chapter 3 we presented a software tool that covers the whole workflow of a TMS experiment. This tool is composed of four dis-

tinct modules, each one addressing a specific aspect of a TMS-based experiment. In particular, the experiment data module manages the patients and the results of the experiments, while the hardware interface module is responsible for the experiment execution and for the interaction with the acquisition equipment. The diagnosis support system is based on an on-line SVM to automatically classify patients based on the MEP responses. The entire platform's data is handled by the data storage module that incorporates semantic web standards to store the data in four distinct repositories for faster access, sensitive data isolation and easier data sharing. Moreover, the proposed RDF schema to describe TMS data allows neuroscientists to share with the neuroscience community both single experiments and entire scientific research studies (data sets and results) with the main aim to standardize the method (i.e. the used variables and procedures/protocols) of studying cortical excitability using TMS. The tool was used during a TMS experiment for evaluating objective parameters for the diagnosis of Vascular Dementia in older patients[130].

Future work on the tool will regard enhancing the automatic signal correction and denoising algorithms for more accurate results. Another important enhancement should be the integration of an advanced dynamic feature selection module so that the DSS can use not only the features derived from MEP signals but also the other patient's variables. To achieve an even tighter integration between non TMS data collection procedures and their joint analysis with TMS data, we are currently working on adding to the proposed tool all the tests (neurophysiological, neuropsychological, etc.) that can be performed using a computer, and also on including available modules for automatic analysis of medical images, in particular the segmentation approaches

proposed in [131] and [74] and the MRI lesion classification module proposed in [132], to reduce the time and effort that the MRI-related variables calculation requires. Finally, a personal health record management system, such as the one in [133], is under development to make the system's patient records globally available.

In Chapter 4 we presented an automatic skeletal bone age assessment method (and the related tool) that implements the TW2 method for EMROI classification and employs Hidden Markov Models for modeling the different stages of development of the bones. The tool can be downloaded at `http://perceive.dieei.unict.it/`. The system was tested and compared against existing state-of-the-art methods and outperformed all of them achieving a correct detection rate of more than 95%, when single EMROIs are concerned. Although the system's performance are very good in terms of accuracy there is much space for improvement and in particular in the preprocessing step, where the DoG filter could be replaced by more sensitive and precise image segmentation methods, such as Markov Random Fields [134]. Moreover, data-mining approaches [135, 136] combined with multimedia retrieval applications [137] could be integrated in order to make more efficient and precise the training process.

In Chapter 5 we have presented BioCloud, an open source, cloud-based platform that assists life science researchers in knowledge discovery. In particular, by integrating text mining methods on scientific documents found in PubMed and Oxford Journal with high-throughput microarrays, BioCloud is able to identify possible gene-gene, gene-disease, protein-disease and protein-protein associations that may be involved in biological processes. A natural language processing module was included in order to find gene-disease relations

in the examined documents. Given that knowledge discovery applications often produce invalid results, the obtained associations were validated against high-throughput annotated biological data. The results shown how parsing full text papers and porting an application on the cloud with minimal effort in terms of programming, increases the efficiency of the platform giving it a net advantage against locally executed processes.

In the near future, we aim at publishing the tool as a free SaaS service to make it available for other users who may want to integrate it in their platform. Since this tool is written in the Java programming language, a transition in the SaaS service model will be quite indolent. Future development will be focused on implementing and optimizing the program for cloud execution. These modifications will regard the complete parallelization of the tool in order to reduce drastrically the time required for processing, even when the number of documents is much larger than the one herein used. Multimedia retrieval methods [138] and image processing [139, 140] could be used for extracting semantic information from images contained in the scientific papers under examination of the application in order to increase the number of the discovered associations.

BIBLIOGRAPHY

[1] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: a survey," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 157–177, 2005.

[2] S. J. Wang, B. Middleton, L. A. Prosser, C. G. Bardon, C. D. Spurr, P. J. Carchidi, A. F. Kittler, R. C. Goldszer, D. G. Fairchild, A. J. Sussman, *et al.*, "A cost-benefit analysis of electronic medical records in primary care," *The American journal of medicine*, vol. 114, no. 5, pp. 397–403, 2003.

[3] X. Zhou, H. Han, I. Chankai, A. A. Prestrud, and A. D. Brooks, "Converting semi-structured clinical medical records into information and knowledge," in *Data Engineering Workshops, 2005. 21st International Conference on*, pp. 1162–1162, IEEE, 2005.

[4] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," in *Proceedings of the 2006 ACM symposium on Applied computing*, pp. 235–239, ACM, 2006.

[5] H. Cunningham, "Gate, a general architecture for text engineering," *Computers and the Humanities*, vol. 36, no. 2, pp. 223–254, 2002.

[6] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[7] Q. Li and Y.-F. B. Wu, "Identifying important concepts from medical documents," *Journal of biomedical informatics*, vol. 39, no. 6, pp. 668–679, 2006.

[8] K. J. Mitchell, M. J. Becich, J. J. Berman, W. W. Chapman, J. Gilbertson, D. Gupta, J. Harrison, E. Legowski, and R. S. Crowley, "Implementation and evaluation of a negation tagger in a pipeline-based system for information extraction from pathology reports," *Medinfo*, vol. 2004, pp. 663–7, 2004.

[9] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries," *Journal of biomedical informatics*, vol. 34, no. 5, pp. 301–310, 2001.

[10] A. Lenci, R. Bartolini, N. Calzolari, A. Agua, S. Busemann, E. Cartier, K. Chevreau, and J. Coch, "Multilingual summarization by integrating linguistic resources in the mlis-musi project.," in *LREC*, vol. 2, pp. 1464–1471, 2002.

[11] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Mashuichi, and K. Ohe, "Text2table: medical text summarization system based on named entity recognition and modality identification,"

in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 185–192, Association for Computational Linguistics, 2009.

[12] D. B. Johnson, Q. Zou, J. D. Dionisio, V. Z. Liu, and W. W. Chu, "Modeling medical content for automated summarization," *Annals of the New York Academy of Sciences*, vol. 980, no. 1, pp. 247–258, 2002.

[13] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[14] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program.," in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.

[15] R. Chen, J. Classen, C. Gerloff, P. Celnik, E. Wassermann, M. Hallett, and L. Cohen, "Depression of motor cortex excitability by low-frequency transcranial magnetic stimulation," *Neurology*, vol. 48, no. 5, pp. 1398–1403, 1997.

[16] M. W. O'Dell, C.-C. D. Lin, and V. Harrison, "Stroke rehabilitation: strategies to enhance motor recovery," *Annual review of medicine*, vol. 60, pp. 55–68, 2009.

[17] L. A. Wheaton, F. Villagra, D. F. Hanley, R. F. Macko, and L. W. Forrester, "Reliability of tms motor evoked potentials in

quadriceps of subjects with chronic hemiparesis after stroke," *Journal of the neurological sciences*, vol. 276, no. 1, pp. 115–117, 2009.

[18] R. Chen, D. Cros, A. Curra, V. Di Lazzaro, J.-P. Lefaucheur, M. R. Magistris, K. Mills, K. M. Rösler, W. J. Triggs, Y. Ugawa, *et al.*, "The clinical diagnostic utility of transcranial magnetic stimulation: report of an ifcn committee," *Clinical Neurophysiology*, vol. 119, no. 3, pp. 504–532, 2008.

[19] F. Mori, C. Ljoka, E. Magni, C. Codecà, H. Kusayanagi, F. Monteleone, A. Sancesario, G. Bernardi, G. Koch, C. Foti, *et al.*, "Transcranial magnetic stimulation primes the effects of exercise therapy in multiple sclerosis," *Journal of neurology*, vol. 258, no. 7, pp. 1281–1287, 2011.

[20] R. Traversa, P. Cicinelli, M. Oliveri, M. G. Palmieri, M. M. Filippi, P. Pasqualetti, and P. M. Rossini, "Neurophysiological follow-up of motor cortical output in stroke patients," *Clinical neurophysiology*, vol. 111, no. 9, pp. 1695–1703, 2000.

[21] G. Pennisi, R. Ferri, M. Cantone, G. Lanza, M. Pennisi, L. Vinciguerra, G. Malaguarnera, and R. Bella, "A review of transcranial magnetic stimulation in vascular dementia," *Dementia and geriatric cognitive disorders*, vol. 31, no. 1, pp. 71–80, 2011.

[22] R. Cantello, R. Tarletti, and C. Civardi, "Transcranial magnetic stimulation and parkinsons disease," *Brain Research Reviews*, vol. 38, no. 3, pp. 309–327, 2002.

[23] M. Kobayashi and A. Pascual-Leone, "Transcranial magnetic stimulation in neurology," *The Lancet Neurology*, vol. 2, no. 3, pp. 145–156, 2003.

[24] F. Fregni and A. Pascual-Leone, "Transcranial magnetic stimulation for the treatment of depression in neurologic disorders," *Current psychiatry reports*, vol. 7, no. 5, pp. 381–390, 2005.

[25] G. Chibbaro, M. Daniele, G. Alagona, C. Di Pasquale, M. Cannavò, V. Rapisarda, R. Bella, and G. Pennisi, "Repetitive transcranial magnetic stimulation in schizophrenic patients reporting auditory hallucinations," *Neuroscience letters*, vol. 383, no. 1, pp. 54–57, 2005.

[26] W. Hu, X. Wang, Y. Yang, D. Liang, and F. Zhao, "[design of a half solenoid coil for optimization of magnetic focusing in transcranial magnetic stimulation].," *Sheng wu yi xue gong cheng xue za zhi= Journal of biomedical engineering= Shengwu yixue gongchengxue zazhi*, vol. 24, no. 4, pp. 910–913, 2007.

[27] P. M. Rossini and S. Rossi, "Transcranial magnetic stimulation diagnostic, therapeutic, and research potential," *Neurology*, vol. 68, no. 7, pp. 484–488, 2007.

[28] T. Kujirai, M. Caramia, J. C. Rothwell, B. Day, P. Thompson, A. Ferbert, S. Wroe, P. Asselman, and C. D. Marsden, "Corticocortical inhibition in human motor cortex.," *The Journal of physiology*, vol. 471, no. 1, pp. 501–519, 1993.

[29] R. Chen, A. Tam, C. Bütefisch, B. Corwell, U. Ziemann, J. C. Rothwell, and L. G. Cohen, "Intracortical inhibition and facili-

tation in different representations of the human motor cortex," *Journal of Neurophysiology*, vol. 80, no. 6, pp. 2870–2881, 1998.

[30] A. Kaelin-Lang and L. Cohen, "Enhancing the quality of studies using transcranial magnetic and electrical stimulation with a new computer-controlled system," *Journal of neuroscience methods*, vol. 102, no. 1, pp. 81–89, 2000.

[31] A. Faro, D. Giordano, I. Kavasidis, C. Pino, C. Spampinato, M. Cantone, G. Lanza, and M. Pennisi, "An interactive tool for customizing clinical transacranial magnetic stimulation (tms) experiments," in *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, pp. 200–203, Springer, 2010.

[32] N. C. Fox and P. A. Freeborough, "Brain atrophy progression measured from registered serial mri: validation and application to alzheimer's disease," *Journal of Magnetic Resonance Imaging*, vol. 7, no. 6, pp. 1069–1075, 1997.

[33] K. A. Jellinger, "The pathology of ischemic-vascular dementia: an update," *Journal of the neurological sciences*, vol. 203, pp. 153–157, 2002.

[34] G. Alagona, R. Ferri, G. Pennisi, A. Carnemolla, T. Maci, E. Domina, A. M. de Noordhout, and R. Bella, "Motor cortex excitability in alzheimer's disease and in subcortical ischemic vascular dementia," *Neuroscience letters*, vol. 362, no. 2, pp. 95–98, 2004.

[35] K. M. Langa, N. L. Foster, and E. B. Larson, "Mixed dementia: emerging concepts and therapeutic implications," *Jama*, vol. 292, no. 23, pp. 2901–2908, 2004.

[36] C. Lippa, T. Smith, and J. M. Swearer, "Alzheimer's disease and lewy body disease: a comparative clinicopathological study," *Annals of neurology*, vol. 35, no. 1, pp. 81–88, 1994.

[37] E. J. Byrne, G. Lennox, J. Lowe, and R. B. Godwin-Austen, "Diffuse lewy body disease: clinical features in 15 cases.," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 52, no. 6, pp. 709–717, 1989.

[38] A. Faro, D. Giordano, M. Pennisi, G. Scarciofalo, C. Spampinato, and F. Tramontana, "Transcranial magnetic stimulation (tms) to evaluate and classify mental diseases using neural networks," in *Artificial Intelligence in Medicine*, pp. 310–314, Springer, 2005.

[39] A. Faro, D. Giordano, M. Pennisi, G. Scarciofalo, C. Spampinato, and F. Tramontana, "A fuzzy model and tool to analyze sivd diseases using tms," *International Journal of Signal Processing*, vol. 2, no. 1, 2006.

[40] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[41] N. A. Syed, S. Huan, L. Kah, and K. Sung, "Incremental learning with support vector machines," 1999.

[42] S. Rüping, "Incremental learning with support vector machines," in *icdm*, p. 641, IEEE, 2001.

[43] I. W. Tsang, J. T. Kwok, and P.-M. Cheung, "Core vector machines: Fast svm training on very large data sets," in *Journal of Machine Learning Research*, pp. 363–392, 2005.

[44] I. W. Tsang, A. Kocsor, and J. T. Kwok, "Simpler core vector machines with enclosing balls," in *Proceedings of the 24th international conference on Machine learning*, pp. 911–918, ACM, 2007.

[45] J. Zheng, F. Shen, H. Fan, and J. Zhao, "An online incremental learning support vector machine for large-scale data," *Neural Computing and Applications*, vol. 22, no. 5, pp. 1023–1035, 2013.

[46] A. Rosenfeld, "Picture processing by computer," *ACM Computing Surveys (CSUR)*, vol. 1, no. 3, pp. 147–176, 1969.

[47] A. Todd-Pokropek, "Image processing in nuclear medicine," *Nuclear Science, IEEE Transactions on*, vol. 27, no. 3, pp. 1080–1094, 1980.

[48] B. F. Hutton and M. Braun, "Software for image registration: algorithms, accuracy, efficacy," in *Seminars in nuclear medicine*, vol. 33, pp. 180–192, Elsevier, 2003.

[49] P. Jannin, J. M. Fitzpatrick, D. Hawkes, X. Pennec, R. Shahidi, M. Vannier, *et al.*, "Validation of medical image processing in image-guided therapy.," *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1445–9, 2002.

[50] L. K. Midyett, W. V. Moore, and J. D. Jacobson, "Are pubertal changes in girls before age 8 benign?," *Pediatrics*, vol. 111, no. 1, pp. 47–51, 2003.

[51] V. Gilsanz and O. Ratib, *Hand bone age: a digital atlas of skeletal maturity*. Springer, 2007.

[52] A. Schmeling, A. Olze, W. Reisinger, F. Rösing, and G. Geserick, "Forensic age diagnostics of living individuals in criminal proceedings," *HOMO-Journal of Comparative Human Biology*, vol. 54, no. 2, pp. 162–169, 2003.

[53] A. Hjern, M. Brendler-Lindqvist, and M. Norredam, "Age assessment of young asylum seekers," *Acta Paediatrica*, 2012.

[54] W. W. Greulich and S. I. Pyle, "Radiographic atlas of skeletal development of the hand and wrist," *The American Journal of the Medical Sciences*, vol. 238, no. 3, p. 393, 1959.

[55] J. Tanner and R. Whitehouse, "Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty.," *Archives of Disease in Childhood*, vol. 51, no. 3, pp. 170–179, 1976.

[56] J. M. Tanner, D. Oshman, G. Lindgren, J. Grunbaum, R. Elouki, and D. Labarthe, "Reliability and validity of computer-assisted estimates of tanner-whitehouse skeletal maturity (casas): comparison with the manual method," *Hormone Research in Paediatrics*, vol. 42, no. 6, pp. 288–294, 1994.

[57] G. Milner, R. Levick, and R. Kay, "Assessment of bone age: a comparison of the greulich and pyle, and the tanner and whitehouse methods," *Clinical radiology*, vol. 37, no. 2, pp. 119–121, 1986.

[58] R. Groell, F. Lindbichler, T. Riepl, L. Gherra, A. Roposch, and R. Fotter, "The reliability of bone age determination in central european children using the greulich and pyle method.," *British journal of radiology*, vol. 72, no. 857, pp. 461–464, 1999.

[59] D. King, D. Steventon, M. O'Sullivan, A. Cook, V. Hornsby, I. Jefferson, and P. King, "Reproducibility of bone ages when performed by radiology registrars: an audit of tanner and whitehouse ii versus greulich and pyle methods," *British journal of radiology*, vol. 67, no. 801, pp. 848–851, 1994.

[60] R. Bull, P. Edwards, P. Kemp, S. Fry, and I. Hughes, "Bone age assessment: a large scale comparison of the greulich and pyle, and tanner and whitehouse (tw2) methods," *Archives of disease in childhood*, vol. 81, no. 2, pp. 172–173, 1999.

[61] M. Niemeijer, B. van Ginneken, C. A. Maas, F. J. Beek, and M. A. Viergever, "Assessing the skeletal age from a hand radiograph: automating the tanner- whitehouse method," in *Proceedings of SPIE*, vol. 5032, pp. 1197–1205, 2003.

[62] I. Kavasidis, C. Pino, and E. Sicurezza, "Automatic skeletal bone age assessment: State of the art and future directions," in *Biomedical Engineering/765: Telehealth/766: Assistive Technologies*, ACTA Press, 2012.

[63] D. Franklin, "Forensic age estimation in human skeletal remains: current concepts and future directions," *Legal Medicine*, vol. 12, no. 1, pp. 1–7, 2010.

[64] P. Thangam, T. Mahendiran, and K. Thanushkodi, "Skeletal bone age assessment–research directions," *Journal of Engineering Science and Technology Review*, vol. 5, no. 1, pp. 90–96, 2012.

[65] E. Pietka, M. F. McNitt-Gray, M. Kuo, and H. Huang, "Computer-assisted phalangeal analysis in skeletal age assessment," *Medical Imaging, IEEE Transactions on*, vol. 10, no. 4, pp. 616–620, 1991.

[66] E. Pietka, L. Kaabi, H. Huang, and M. Kuo, "Feature extraction for bone age determination," *Radiology*, vol. 177, 1990.

[67] A. Zhang, A. Gertych, and B. J. Liu, "Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones," *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, vol. 31, no. 4-5, p. 299, 2007.

[68] C.-W. Hsieh, T.-C. Liu, T.-L. Jong, and C.-M. Tiu, "A fuzzy-based growth model with principle component analysis selection for carpal bone-age assessment," *Medical and Biological Engineering and Computing*, vol. 48, no. 6, pp. 579–588, 2010.

[69] C.-W. Hsieh, T.-L. Jong, and C.-M. Tiu, "Bone age estimation based on phalanx information with fuzzy constrain of carpals,"

Medical and Biological Engineering and Computing, vol. 45, no. 3, pp. 283–295, 2007.

[70] A. Gertych, A. Zhang, J. Sayre, S. Pospiech-Kurkowska, and H. Huang, "Bone age assessment of children using a digital hand atlas," *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, vol. 31, no. 4-5, p. 322, 2007.

[71] S. Mahmoodi, B. Sharif, E. Chester, J. Owen, and R. Lee, "Automated vision system for skeletal age assessment using knowledge based techniques," in *Image Processing and Its Applications, 1997., Sixth International Conference on*, vol. 2, pp. 809–813, IET, 1997.

[72] F. Vogelsang, M. Kohnen, H. Schneider, F. Weiler, M. W. Kilbinger, B. B. Wein, and R. W. Guenther, "Skeletal maturity determination from hand radiograph by model based analysis," in *SPIE proceedings series*, pp. vol1–294, Society of Photo-Optical Instrumentation Engineers, 2000.

[73] D. Giordano, C. Spampinato, G. Scarciofalo, and R. Leonardi, "Automatic skeletal bone age assessment by integrating emroi and croi processing," in *Medical Measurements and Applications, 2009. MeMeA 2009. IEEE International Workshop on*, pp. 141–145, IEEE, 2009.

[74] D. Giordano, C. Spampinato, G. Scarciofalo, and R. Leonardi, "An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones," *In-*

strumentation and Measurement, IEEE Transactions on, vol. 59, no. 10, pp. 2539–2553, 2010.

[75] H. Y. Chai, L. K. Wee, T. T. Swee, and S.-H. Salleh, "Adaptive crossed reconstructed (acr) k-mean clustering segmentation for computer-aided bone age assessment system," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 5, no. 3, pp. 628–635, 2011.

[76] M. Harmsen, B. Fischer, H. Schramm, and T. M. Deserno, "Support vector machine classification using correlation prototypes for bone age assessment," *Bildverarbeitung für die Medizin 2012*, pp. 434–439, 2012.

[77] D. Haak, J. Yu, H. Simon, H. Schramm, T. Seidl, and T. M. Deserno, "Bone age assessment using support vector regression with smart class mapping," in *SPIE Medical Imaging*, pp. 86700A–86700A, International Society for Optics and Photonics, 2013.

[78] H. H. Thodberg, S. Kreiborg, A. Juul, and K. D. Pedersen, "The bonexpert method for automated determination of skeletal maturity," *Medical Imaging, IEEE Transactions on*, vol. 28, no. 1, pp. 52–66, 2009.

[79] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H. Huang, and V. Gilsanz, "Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal roi extraction," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 8, pp. 715–729, 2001.

[80] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[81] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[82] E. Pietka, S. Pospiech-Kurkowska, A. Gertych, and F. Cao, "Integration of computer assisted bone age assessment with clinical pacs," *Computerized medical imaging and graphics*, vol. 27, no. 2, pp. 217–228, 2003.

[83] A. Özgür, T. Vu, G. Erkan, and D. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277–i285, 2008.

[84] S. Ananiadou, D. B. Kell, and J.-i. Tsujii, "Text mining and its potential applications in systems biology.," *Trends in biotechnology*, vol. 24, pp. 571–579, Dec. 2006.

[85] Y. Liu, S. Navathe, J. Civera, V. Dasigi, A. Ram, B. Ciliax, and R. Dingledine, "Text mining biomedical literature for discovering gene-to-gene relationships: a comparative study of algorithms," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 2, no. 1, pp. 62–76, 2005.

[86] C. Von Mering, L. Jensen, B. Snel, S. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. Huynen, and P. Bork, "String: known and predicted protein–protein associations, integrated

and transferred across organisms," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D433–D437, 2005.

[87] A. Faro, D. Giordano, and C. Spampinato, "Combining literature text mining with microarray data: advances for system biology modeling," *Brief. Bioinformatics*, vol. 13, pp. 61–82, Jan 2012.

[88] S. Ekins, Y. Nikolsky, A. Bugrim, and et al., "Pathway mapping tools for analysis of high content data.," *Methods in molecular biology (Clifton, N.J.)*, vol. 356, pp. 319–350, 2007.

[89] J. Wu, X. Mao, T. Cai, J. Luo, and L. Wei, "KOBAS server: a web-based platform for automated annotation and pathway identification," *Nucleic Acids Res.*, vol. 34, pp. W720–724, Jul 2006.

[90] R. Frijters, B. Heupers, P. van Beek, and et al., "CoPub: a literature-based keyword enrichment tool for microarray data analysis," *Nucleic Acids Res.*, vol. 36, pp. W406–410, Jul 2008.

[91] L. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, and Y. Moreau, "Endeavour update: a web resource for gene prioritization in multiple species," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W377–W384, 2008.

[92] C. Perez-Iratxeta, M. Wjst, P. Bork, and M. Andrade, "G2d: a tool for mining genes associated with disease," *BMC genetics*, vol. 6, no. 1, p. 45, 2005.

[93] F. Azuaje, J. Dopazo, and J. Wiley, *Data analysis and visualization in genomics and proteomics*. Wiley Online Library, 2005.

[94] P. Mell and T. Grance, "The nist definition of cloud computing (draft)," *NIST special publication*, vol. 800, p. 145, 2011.

[95] A. Bateman and M. Wood, "Cloud computing," *Bioinformatics*, vol. 25, no. 12, pp. 1475–1475, 2009.

[96] A. Hey, S. Tansley, and K. Tolle, *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research Redmond, WA, 2009.

[97] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[98] M. Schatz, "Cloudburst: highly sensitive read mapping with mapreduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[99] J. Shendure and H. Ji, "Next-generation dna sequencing," *Nature biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[100] A. Matsunaga, M. Tsugawa, and J. Fortes, "Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications," in *eScience, 2008. eScience'08. IEEE Fourth International Conference on*, pp. 222–229, IEEE, 2008.

[101] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. Madden, "Ncbi blast: a better web interface," *Nucleic acids research*, vol. 36, no. suppl 2, pp. W5–W9, 2008.

[102] L. Stein *et al.*, "The case for cloud computing in genome informatics," *Genome Biol*, vol. 11, no. 5, p. 207, 2010.

[103] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," *Briefings in Bioinformatics*, vol. 11, no. 5, pp. 473–483, 2010.

[104] R. Grossman and Y. Gu, "Data mining using high performance data clouds: experimental studies using sector and sphere," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 920–927, ACM, 2008.

[105] D. Borthakur, "The hadoop distributed file system: Architecture and design," *Hadoop Project Website*, vol. 11, p. 21, 2007.

[106] A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy, "Hive: a warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1626–1629, 2009.

[107] H. Liu and D. Orban, "Cloud mapreduce: a mapreduce implementation on top of a cloud operating system," in *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*, pp. 464–474, IEEE, 2011.

[108] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[109] P. Watson, D. Leahy, H. Hiden, S. Woodman, and J. BerryLiu, "An azure science cloud for drug discovery," *Microsoft External Research Symposium*, 2009.

[110] X. Qiu, J. Ekanayake, S. Beason, T. Gunarathne, G. Fox, R. Barga, and D. Gannon, "Cloud technologies for bioinformatics applications," in *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers*, p. 6, ACM, 2009.

[111] F. Nazareno, K. Lee, and W. Cho, "Mining molecular interactions from scientific literature using cloud computing," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pp. 864–865, IEEE, 2010.

[112] J. Lin and C. Dyer, "Data-intensive text processing with mapreduce," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–177, 2010.

[113] B. Langmead, M. Schatz, J. Lin, M. Pop, and S. Salzberg, "Searching for snps with cloud computing," *Genome Biol*, vol. 10, no. 11, p. R134, 2009.

[114] M. Schatz, D. Sommer, D. Kelley, and M. Pop, "De novo assembly of large genomes using cloud computing," in *CSHL Biology of Genomes conference*, 2010.

[115] B. Langmead, K. Hansen, and J. Leek, "Cloud-scale rna-sequencing differential expression analysis with myrna," *Genome Biol*, vol. 11, no. 8, p. R83, 2010.

[116] M. Karim, A. Bari, B. Jeong, and H. Choi, "Cloud technology for mining association rules in microarray gene expression datasets,"

[117] J. Delmerico, N. Byrnes, A. Bruno, M. Jones, S. Gallo, and V. Chaudhary, "Comparing the performance of clusters, hadoop, and active disks on microarray correlation computations," in *High Performance Computing (HiPC), 2009 International Conference on*, pp. 378–387, IEEE, 2009.

[118] T. Nguyen, W. Shi, and D. Ruden, "Cloudaligner: A fast and full-featured mapreduce based tool for sequence mapping," *BMC research notes*, vol. 4, no. 1, p. 171, 2011.

[119] V. Fusaro, P. Patil, E. Gafni, D. Wall, and P. Tonellato, "Biomedical cloud computing with amazon web services," *PLoS computational biology*, vol. 7, no. 8, p. e1002147, 2011.

[120] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, K. Nelson, *et al.*, "Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC bioinformatics*, vol. 13, no. 1, p. 42, 2012.

[121] S. Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D. Riley, C. Arze, J. White, O. White, and W. Fricke, "Clovr: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC bioinformatics*, vol. 12, no. 1, p. 356, 2011.

[122] S. Angiuoli, J. White, M. Matalka, O. White, and W. Fricke, "Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing," *PloS one*, vol. 6, no. 10, p. e26624, 2011.

[123] J. Dudley, Y. Pouliot, R. Chen, A. Morgan, and A. Butte, "Translational bioinformatics in the cloud: an affordable alternative," *Genome medicine*, vol. 2, no. 8, p. 51, 2010.

[124] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters, *Text Processing with GATE (Version 6)*. 2011.

[125] B. Settles, "Abner: an open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[126] B. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. Baseler, H. Lane, R. Lempicki, *et al.*, "David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W169–W175, 2007.

[127] A. I. Saeed, V. Sharov, J. White, and et al., "TM4: a free, open-source system for microarray data management and analysis," *BioTechniques*, vol. 34, pp. 374–378, Feb 2003.

[128] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of Computational Biology*, vol. 6, no. 3/4, pp. 281–297, 1999.

[129] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007.

[130] C. Spampinato, E. Aguglia, C. Concerto, M. Pennisi, G. Lanza, R. Bella, M. Cantone, G. Pennisi, I. Kavasidis, and D. Giordano, "Transcranial magnetic stimulation in the assessment of motor cortex excitability and treatment of drug-resistant major depression," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, no. 3, pp. 391–403, 2013.

[131] D. Giordano, R. Leonardi, F. Maiorana, G. Scarciofalo, and C. Spampinato, "Epiphysis and metaphysis extraction and classification by adaptive thresholding and dog filtering for automated skeletal bone age analysis," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 6551–6556, IEEE, 2007.

[132] A. Faro, D. Giordano, C. Spampinato, and M. Pennisi, "Statistical texture analysis of mri images to classify patients affected by multiple sclerosis," in *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, pp. 272–275, Springer, 2010.

[133] A. Faro, D. Giordano, I. Kavasidis, and C. Spampinato, "A

web 2.0 telemedicine system integrating tv-centric services and personal health records," in *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on*, pp. 1–4, IEEE, 2010.

[134] D. Anguelov, B. Taskarf, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng, "Discriminative learning of markov random fields for segmentation of 3d scan data," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 169–176, IEEE, 2005.

[135] A. Zhang, F. Cao, E. Pietka, B. J. Liu, and H. Huang, "Data mining for average images in a digital hand atlas," in *Proceedings of SPIE*, vol. 5371, pp. 251–258, 2004.

[136] C. Faloutsos and K.-I. Lin, *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*, vol. 24. ACM, 1995.

[137] D. Giordano, I. Kavasidis, C. Pino, and C. , "A semantic-based and adaptive architecture for automatic multimedia retrieval composition," in *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pp. 181–186, IEEE, 2011.

[138] D. Giordano, I. Kavasidis, C. Pino, and C. Spampinato, "A semantic-based and adaptive architecture for automatic multimedia retrieval composition," in *CBMI 2011*, pp. 181 –186, june 2011.

[139] D. Giordano, R. Leonardi, F. Maiorana, G. Scarciofalo, and C. Spampinato, "Epiphysis and metaphysis extraction and clas-

sification by adaptive thresholding and dog filtering for automated skeletal bone age analysis," in *EMBS 2007*, pp. 6551 – 6556, aug. 2007.

[140] D. Giordano, C. Spampinato, G. Scarciofalo, and R. Leonardi, "An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones.," *IEEE T. Instrumentation and Measurement*, vol. 59, no. 10, pp. 2539–2553, 2010.