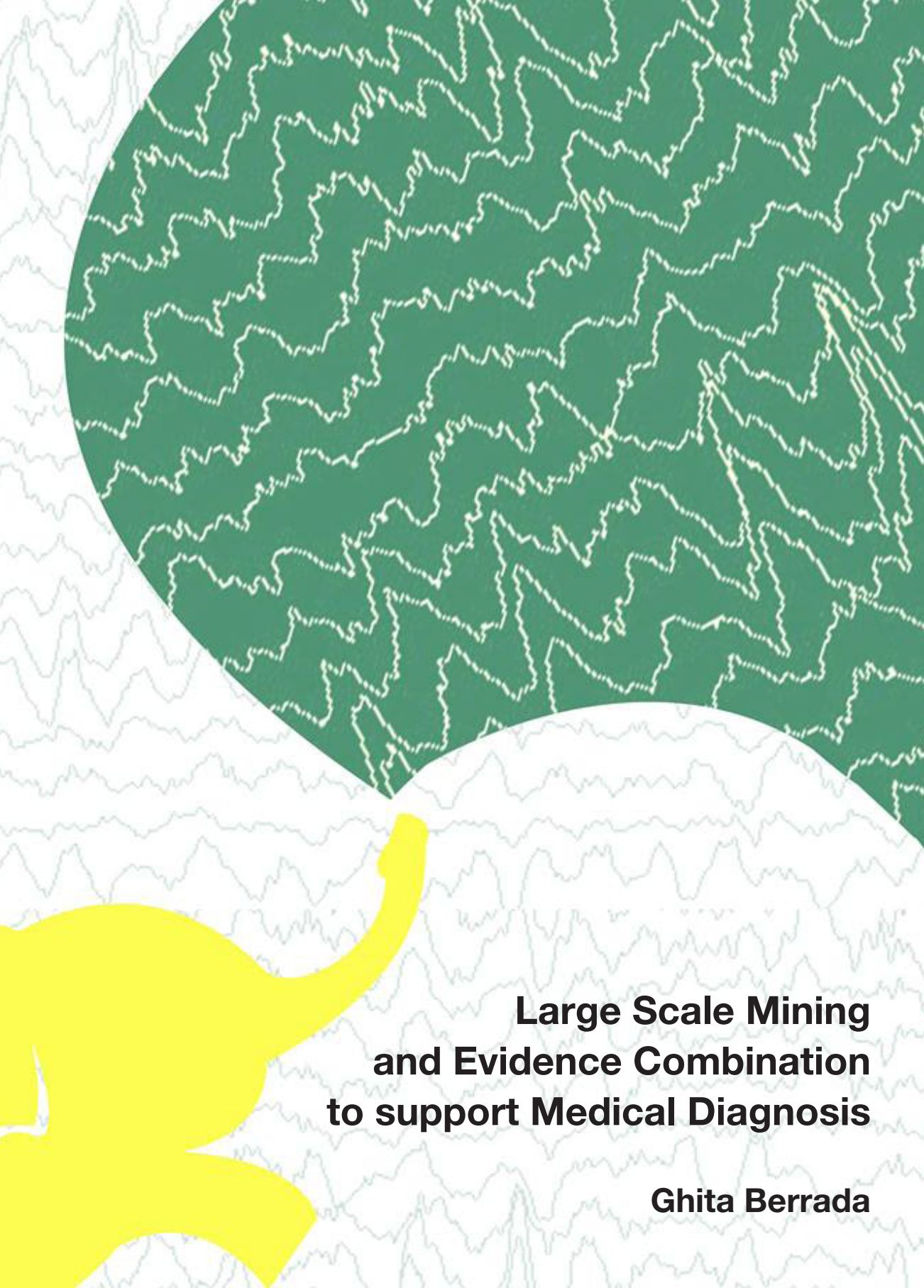


Large Scale Mining and Evidence Combination to support Medical Diagnosis

Ghita Berrada

Large scale mining and evidence combination to support medical diagnosis

Ghita Berrada

Graduation committee:

Chairman:	Prof. dr. Peter M. G. Apers
Promoter:	Prof. dr. Peter M. G. Apers
Assistant promoter:	Dr. ir. Maurice van Keulen

Members:

Prof. dr. ir. Michel J. A.M van Putten	University of Twente
Dr. ir. Bennie ten Haken	University of Twente
Prof. Ian T. Nabney	Aston University (Birmingham,UK)
Prof. Guy de Tré	Ghent University (Ghent, Belgium)

The research presented in this thesis was funded as part of the ViP Brain Networks project supported by the Dutch Ministry of Economic Affairs, Agriculture and Innovation, province Overijssel and province Gelderland.

The research was performed at the University of Twente, at the Faculty of Science and Technology (TNW) (Clinical Neurophysiology (CNPH) and Neuroimaging (NIM) groups) as well as in the Database Group (DB) at the Faculty of Electrical Engineering, Mathematics and Computer Science (EWI).

CTIT

CTIT Ph.D.-thesis Series No. 14-331
Centre for Telematics and Information Technology
University of Twente
P.O. Box 217, NL – 7500 AE Enschede

ISSN 1381-3617 (CTIT Ph.D. thesis Series No. 14-331)
ISBN 978-90-365-3825-1
DOI 10.3990/1.9789036538251
http://dx.doi.org/10.3990/1.9789036538251

Printed by: Optima Grafische Communicatie, Rotterdam
Cover design: Ariane Hofmann-Maniyar

LARGE SCALE MINING AND EVIDENCE COMBINATION TO SUPPORT MEDICAL DIAGNOSIS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
Prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties,
in het openbaar te verdedigen
op vrijdag 16 januari 2015 om 12.45 uur

door

Ghita Berrada

geboren op 2 january 1984
te Rabat, Morocco

Dit proefschrift is goedgekeurd door:
Prof. dr. ir. Peter M. G Apers(promotor)
Dr. ir. Maurice van Keulen(assistent-promotor)

To my parents and sisters, for their unflinching love and support
In loving memory of my grandparents

Acknowledgments

The PhD journey is finally coming to an end. What a journey it has been! Meandering, unpredictable, chaotic but all the same exhilarating, colorful and memorable. I am thankful for the journey because it not only helped me grow as a scientist but also as a human being. And I am thankful to all the people who have been by my side for part of the journey or for it all: family, friends, colleagues, acquaintances. Without them all, I wouldn't be standing where I am today, I wouldn't be who I am today and I wouldn't have experienced as many things as I have since the journey started.

First and foremost, I want to thank all my supervisors and promotors through the years: Michel van Putten, Ander de Keijzer, Christian Beckmann, Maurice van Keulen and Peter Apers. Without your guidance and encouragements, none of this would have been possible. I am also grateful to you all for teaching me what being a researcher truly entails. I am particularly grateful to Michel van Putten for providing me the quality data without which this thesis would not have been possible. A special thank you to Christian Beckmann, Maurice van Keulen and Peter Apers for stepping in when I hit rough patches during my PhD and when my continuing and finishing my PhD was by no means guaranteed. I would also like to extend a very special, heartfelt thank you to Maurice van Keulen. Thank you for believing in me and guiding me until the end. Thank you as well for being so understanding and caring at all times: I am sure it must not have been easy all the time.

I would also like to thank all my committee members Guy de Tré, Ian Nabney, Bennie ten Haken and Michel van Putten for their time, interest in my thesis and insightful comments. And also thank you for proofreading the thesis so thoroughly and spotting all the typos for me to fix.

I consider myself extremely lucky to have been part of the databases (DB) group, if only temporarily. More than colleagues, all of you have been friends

who have made my time in the office thoroughly enjoyable and memorable. Thank you Mohammad, Lei, Zhemin, Brend, Robin, Victor, Djoerd, Rezwan, Juan and Jan.
A special mention to Mena for giving me precious advice at several points in my PhD and particularly during the thesis writing process (as well as for being one of the few people I could discuss tennis with).
I am also grateful to Brend for helping translate my thesis abstract into Dutch and enduring my "Yoda-Dutch", Djoerd for giving me access to the Hadoop servers while I was still in the Technical Medicine department and to Robin for his help with Hadoop. I must also thank Jan for all his help with the servers, even and particularly when they went down due to some of my experiments.
A special mention to Ida, the soul and pillar of the DB group (until a few weeks ago). To Ida, nothing is impossible. Because you were there, I just knew that I needn't worry whatever came my way. Thank you for everything Ida and "success" in your new group.

I was also privileged to have been part of the Clinical Neurophysiology and NIM groups (Technical Medicine department, TNW) and would like to thank all the people in both groups for their kindness and support and for making me consider my thesis subject under a different point of view than the one I am used to.
Special thanks to the secretaries of both groups, Jolanda, Tanja, Esmeralda, Cindy, Claudia, Daniëlle for helping me settle in the Netherlands and sort the diverse administrative issues that came my way. Thank you as well for all the fun discussions and precious advice on various life matters.
A very big thank you to Esther, Chin, Shaun, Bas-Jan, Martijn, Michiel, Elmer, Marleen and Cecil for all the discussions, advice, laughter, ideas and tea times. And thank you for helping me with moving houses countless times as well as for introducing me to the joys of ice-skating.

I am grateful to all my friends in particular Raje, Ming, KissCool (aka Philippe Bayle), mando (aka Marc-Olivier Buob) for all the support and advice I have received from them all these years. And thank you Ariane for helping with the thesis cover design.

I also have to thank all the IND agents who have, throughout the years, processed my visa and residence permit applications for making it possible for me to come to the Netherlands for my PhD and stay until the end of my PhD without too much trouble.

Last but not least, a thought for my family. I have always felt lucky to have been born with you, Mum and Dad, as parents and have you both Nadia and Selma as sisters and never have I felt so lucky than during my PhD years. You gave me the motivation to go for it and supported me in all possible ways all through it. Sorry for the rough times and thank you for everything. Thank you in particular, Mum and Dad, for "teach[ing me] to fish" early on.
I am also grateful to all my uncles, aunts and cousins who, through their concern, love, support or more material help, spurred me on, with a particular thought to uncle Ghali, aunt Houria, uncle Abderrahman and Azizi. And a very special thanks to my cousin Abdelkrim who learnt Hadoop along with me, helped me debug code into the wee hours of the night and without whom the experiments in chapter 3 would not have been possible.

And now a new journey starts...
*Départ dans l'affection et le bruit neufs! (Arthur Rimbaud, "Départ", **Illuminations**)*

Ghita Berrada
Enschede, November 2014

Summary

A few days after having been back home from a first ER visit with a diagnosis of benign flu, teenager Rory Staunton dies of sepsis (a potentially fatal whole-body inflammatory response to a very serious infection) [1].
Though misdiagnoses do not always lead to very serious outcomes such as in this case, they are a major and largely overlooked problem. The prevalence of misdiagnoses is estimated to be up to 15% in most areas of medicine ([2]). And a study of physician-reported diagnosis errors ([3]) finds that most cases are due to testing (44%) or clinician assessment errors (32%) and that 28% of the misdiagnoses are major (i.e resulting in death, permanent disability, or near life-threatening event) and 41% moderate (i.e resulting in short-term morbidity, increased length of stay, higher level of care or invasive procedure). [4] estimates missed diagnoses alone account for 40 000 to 80 000 preventable deaths annually in the US. Zebras [1] are very likely to be misdiagnosed with clinicians trained to look for the most common diagnoses first but even common conditions such as pneumonia, asthma or breast cancer are routinely misdiagnosed especially if the symptoms presentation is atypical ([5, 6]).
The misdiagnosis problem is often considered to be an individual clinician's problem. Yet the facts and figures presented earlier rather suggest misdiagnoses to be more of a systemic problem. Part of the problem stems from the accessibility of patient data, in particular patient history that is credited for being the key factor leading to diagnosis in 56% to 82.5% of the cases according to a review several studies on factors contributing to a diagnosis ([7]). Patient data is currently scattered across various locations often using different platforms and data storage standards and is sometimes not accessible because it is not digitized or discarded after real-time use. A McKinsey Global Institute report on the US healthcare system ([8]) estimates that 30% of data that includes medical records, laboratory and surgery reports, is not digitized and that 90% of the data generated by healthcare providers is discarded, for example almost all video feeds from surgery. In this context, it becomes hard for a clinician to

[1] rare diseases or conditions

get a full picture of a patient's condition and make an informed diagnosis. The problem also comes from the sheer amount of data and its complexity: interpreting test data to come up with a diagnosis often requires specialist knowledge, increases clinicians' workload and is error-prone and far from straightforward. And as clinicians are expected to deliver fast and accurate diagnoses based on incomplete and highly uncertain data, they tend to resort to all kinds of cognitive shortcuts and heuristics that, while useful, increase the likelihood of diagnosis error if misapplied (eg. premature closure bias that leads a clinician to focus on only one diagnosis hypothesis too fast or confirmation bias that makes a clinician reinterpret the evidence at hand to support his/her preferred hypothesis and discard any disproving evidence) ([9]).
There is little doubt, based on this, that some support needs to be given clinicians to make the diagnosis process faster and more accurate. Providing a medical data sharing platform is one of the possible solutions to improve the diagnosis process. Two stakeholder groups stand to benefit from such a shared platform: a patients/clinicians group and a researchers/clinicians group. A shared data platform would allow researchers/clinicians access to a (standard) trove of data on which to develop and test (semi)-automated medical data interpretation methods so as to reduce clinicians' workloads and improve their performance. A shared data platform would also make patient data and in particular history fully and easily accessible, which would help clinicians come up with more accurate diagnoses faster and improve patients' quality of life.
This thesis' goal is to come up with a first design of a shared medical data platform, using EEG data as an example of medical data.
There are three main contributions in this thesis:

1. a feasibility study for medical data sharing and processing platform using Hadoop
2. a proposal for a feature-based similarity measure to perform EEG similarity search
3. a model for evidence combination

The first contribution evaluates Hadoop as a potential platform for medical data sharing and processing platform. In the first contribution, we show (Chapter 3) that Hadoop is the technology needed to share data at little expense and effort. In particular, it explains no effort is needed to standardize the existing data formats as long as methods to read and/or visualize them exist and are made available since Hadoop can handle diverse data formats natively. We also demonstrate in the first contribution that Hadoop is a suitable for developing medical data interpretation methods since one of the most computationally

expensive data mining tasks (ie exhaustive search feature selection) can be performed, on the Hadoop platform, on national scale amounts of representative data (i.e EEG data), thus proving the readiness in performance and scalability for medical data interpretation methods. In a sense, the main argument in the first contribution is that the only step needed to start sharing and processing medical data is to start deploying Hadoop in medical institutions and transferring data to the Hadoop platform.

The second contribution is to propose a similarity measure based on features extracted from EEGs so as to retrieve EEGs once stored in the medical data sharing platform through similarity search. Three features in particular are studied: the fractal dimension, the spectral entropy and the high/low frequency ratio. The features chosen for the similarity measure are EEG-specific but the principle of the similarity search methods can be used for other types of data in particular other medical time series.
Because the medical diagnosis process is incremental, uncertain and evidence-based (eg evidence obtained through user feedback or (semi)-automated medical data interpretation methods), our third contribution is an evidence combination model based on the Dempster-Shafer theory that allows us to quantify the uncertainty attached to each diagnosis alternative. This model takes into account the fact that not all sources of evidence are necessarily of equal reliability.
Contributions 1 and 2 are validated experimentally. There was no user study done for contribution 3 (this could be future work) so contribution 3 was validated theoretically through proving various convergence properties.

Samenvatting

Een paar dagen nadat tiener Rory Stauton terugkeerde van de spoedeisende hulp met de diagnose 'simpel griepje' overleed hij aan sepsis: een soms dodelijke ontstekingsreactie van het hele lichaam als reactie op een infectie. [1].

Alhoewel de gevolgen niet altijd even serieus zijn als het gevolg van de diagnose van Rory Stauton, zijn foutive diagnoses een zwaar en vaak genegeerd probleem. De aanwezigheid van foutieve diagnoses wordt geschat op 15% in de meeste medische vakgebieden ([2]). Een studie over diagnostische fouten onder geneeskundigen ([3]) toont dat de meeste gevallen van foutieve diagnose optreden door medische testen (44%) of door een redenatiefout tijdens de diagnose (32%). Diezelfde studie wijst uit dat 28% van de foutieve diagnoses serieuze consequenties hebben (i.e., leiden tot de dood, permanente invaliditeit of een levensbedreigende situatie) en 41% heeft gematigde consequenties (i.e., resulterend in kortedurende ziekte, een langer ziekenhuisverblijf, een hogere verzorgingsgraad of een invasieve ingreep). [4] schat het aantal vermijdbare doden door verkeerde diagnoses tussen de 40 000 en 80 000 in de V.S. Het is zeer waarschijnlijk dat zeldzame ziektes en condities (de zogenaamde 'zebras') verkeerd gediagnosticeerd worden want geneeskundigen zijn getrained in het herkennen van de meest voorkomende condities. Maar zelfs veel voorkomende condities zoals longontsteking, astma en borstkanker worden regelmatig foutief gediagnosticeerd, zeker als de symptomen afwijken van de norm ([5, 6]).

Foutieve diagnoses worden gezien als het probleem van de individuele geneeskundige. Maar de feiten zoals eerder aangegeven suggereren dat foutieve diagnoses een systematisch probleem zijn. Een van de oorzaken van het probleem is de toegankelijkheid van patiëntdata. Toegang tot de medische geschiedenis van de patiënt wordt in het specifiek aangewezen als belangrijk voor de diagnose in 56% tot 82.5% van de gevallen, volgens een review van verscheidene studies over factoren die bijdragen aan een diagnose ([7]). Patiëntdata is momenteel verdeeld over verschillende locaties, platformen en dataformaten, en is soms niet beschikbaar omdat het niet gedigitaliseerd is of weggegooid

wordt na direct gebruik. Een rapport van het McKinsey Global Institute over het zorgsysteem in de V.S. ([8]) schat dat 30% van de medische records, laben operatierapporten niet gedigitaliseerd is en dan 90% van de gegenereerde data, zoals videos gemaakt tijdens operaties, weggegooid wordt. Hierdoor wordt het moeilijk voor de geneeskundige om een volledig beeld te krijgen van de situatie van een patiënt of om een diagnose te stellen. De complexiteit en de hoeveelheid data draagt ook bij aan het probleem: de intepretatie van testresultaten om een diagnose te stellen vereist specialistische kennis, verhoogt de hoeveelheid werk voor de geneeskundige en is foutgevoelig. Omdat verwacht wordt dat geneeskundigen een snelle en accurate diagnose stellen gebaseerd op incomplete en onzekere informatie neigen ze naar het gebruik van vuistregels en heuristieken die, alhoewel zinnig, leiden tot een verhoogde kans op foutieve diagnose als ze verkeerd toegepast worden (bijv., 'premature closure bias' waardoor de geneeskundige te snel focust op één enkele diagnose, of 'confirmation bias' waardoor de geneeskundige bewijsmateriaal herinterpreteerd om zijn of haar voorkeurshypothese te bevestigen en tekenen dat het anders is te negeren) ([9]).

Gebaseerd op deze informatie is er geen twijfel dat er extra ondersteuning gegeven moet worden aan geneeskundigen om tot een snellere en meer acurate diagnose te komen. Het aanbieden van een platform voor het delen van medische data is een van de mogelijke oplossingen om het diagnostische process te verbeteren. Twee groepen hebben belang bij een dergelijk platform: de patiënt/geneeskundige groep, en de de onderzoeker/geneeskundige groep. Een gedeeld platform biedt de onderzoeker/geneeskundige groep toegang tot een (gestandaardiseerde) schat aan informatie waarmee nieuwe methoden voor (semi)-automatische interpretatie van medische data ontwikkeld en getest kunnen worden. Een gedeeld platform helpt de patiënt door de medische geschiedenis volledige en eenvoudig toegankelijk te maken waardoor geneeskundigen een meer accurate diagnose kunnen stellen en de patiënt beter kunnen helpen. Het doel van dit proefschrift is het bepalen van een, aan de hand van EEG als voorbeeld, eerste ontwerp voor een gedeeld platform voor medische data.

De drie hoofdcontributies van dit proefschrift zijn:

1. een haalbaarheidsstudie voor het gebruik van Hadoop als platform voor het delen en verwerken van medische data

2. een voorstel voor een feature-based similarity measure om EEG similarity search te doen

3. een model voor het combineren van bewijsmateriaal

De eerste contributie is de evaluatie van Hadoop als potentieel platform voor het delen en verwerken van medische data (hoofdstuk 3). We tonen dat Hadoop een technologie is die gebruikt kan worden om data te delen met weinig extra moeite. Ook tonen we dat het gebruik van Hadoop geen extra werk en kosten met zich meebrengt om verschillende dataformaten te hanteren. Zolang methoden om de data te lezen en te visualizeren beschikbaar zijn kan Hadoop deze afhandelen. Verder demonstreren we dat Hadoop geschikt is om medische data mee te interpreten door aan te tonen dan een van de computationeel meest vereisende taken (i.e., 'exhaustive search feature selection') uitgevoerd kan worden op het Hadoop platform op medische data van een landelijke schaal. Hiermee tonen we de gebruiksklaarheid en schaalbaarheid van methoden om medische data te interpreteren. Het punt dat gemaakt wordt in deze contributie is dat het uitrollen van Hadoop en het verplaatsen van data naar dit platform de enige stap is die nodig is om het delen en verwerken van medische data te starten.

De tweede contributie is een voorstel voor een feature based similarity measure op EEG data zodat EEGs opgeslagen op het platform teruggevonden kunnen worden op basis van similarity search. Drie features zijn onderzocht: fractal dimension, spectral entropy en high/low frequency ratio. De gekozen features zijn specifiek voor EEG data, maar het principe van similarity search kan gebruikt worden voor andere soorten data, en medische tijdreeksanalyses bij uitstek.

Omdat het diagnoseprocess incrementeel, onzeker en bewijsmateriaalafhankelijk is (bijv., bewijsmateriaal verkregen door user feedback of (semi)-automatische interpretatie) richt de derde contributie zich op een model om bewijsmateriaal te combineren. Dit model is gebaseerd op de Dempster-Shafer theorie die het mogelijk maakt om de onzekerheid van verschillende alternatieve diagnoses uit te drukken. Het voorgestelde model houdt rekening met het feit dat niet alle bronnen van bewijsmateriaal even betrouwbaar zijn.

De eerste twee contributies zijn experimenteel gevalideerd. Er is geen gebruikersonderzoek gedaan voor de derde contributie, deze contributie is gevalideerd door middel van theoretische bewijzen voor diverse convergentie-eigenschappen.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

An article in the Washington Post ([10]) recounts how a patient barely survived a series of misdiagnoses and a test never to be performed in patients with his condition. After months of being alternately told his symptoms (flu-like symptoms, dizziness, headaches, weight gain and liver problems) were due to his weight, fatigue or tension headaches, the patient ends up in the ER where an ordered CT scan shows a cyst in his brain. The ER physician advises him to follow up on this finding with his doctor, performs a spinal tap to rule out meningitis as a cause of the patient's symptoms and discharges him. A neurosurgeon, who had been given the CT to review as it was abnormal, stops the patient in extremis from leaving with bad tidings: the just discovered cyst was increasing the intracranial pressure causing the patient's symptoms and the spinal tap had not only aggravated the problem but also made it potentially fatal so emergency surgery had to be performed to avoid a lethal outcome.
Though misdiagnoses do not always lead to very serious outcomes such as in this case, they are a major and largely overlooked problem. The prevalence of misdiagnoses is estimated to be up to 15% in most areas of medicine ([2]). Moreover, a study of physician-reported diagnosis errors ([3]) finds that most cases are due to testing (44%) or clinician assessment errors (32%) and that 28% of the misdiagnoses are major [1] and 41% moderate [2]. [4] estimates missed diagnoses alone account for 40 000 to 80 000 preventable deaths annually in the US. Rare diseases or conditions (also called zebras) are very likely to be misdiagnosed since clinicians are trained to look for the most common diagnoses

[1] i.e resulting in death, permanent disability, or near life-threatening event

[2] i.e resulting in short-term morbidity, increased length of stay, higher level of care or invasive procedure

first but even common conditions such as pneumonia, asthma or breast cancer are routinely misdiagnosed especially if the symptoms presentation is atypical ([5, 6]).
Part of the problem stems from the accessibility of patient data, in particular patient history that is credited for being the key factor leading to diagnosis in 56% to 82.5% of the cases according to a review several studies on factors contributing to a diagnosis ([7]). Patient data is currently scattered across various locations often using different platforms and data storage standards and is sometimes not accessible because it is not digitized or discarded after real-time use. A McKinsey Global Institute report on the US healthcare system ([8]) estimates that 30% of data that includes medical records, laboratory and surgery reports, is not digitized and that 90% of the data generated by healthcare providers is discarded, for example almost all video feeds from surgery. In this context, it becomes hard for a clinician to get a full picture of a patient's condition and make an informed diagnosis.
The problem also comes from the sheer amount of data and its complexity: interpreting test data to come up with a diagnosis often requires specialist knowledge, increases clinicians' workload and is error-prone and far from straightforward. Ongoing efforts are being made to develop (semi-)automated methods of data interpretation so as to ease the clinicians' task, help them with the diagnosis process and minimize the interpretation time as well as the risk of error. For instance for EEG data-multidimensional time series corresponding to the electrical signals recorded at different locations of the brain scalp, for further details see Chapter 2-, such methods include:

- [11] that assesses the existence of brain injury/asphyxia and its degree by computing the cepstral distance between the EEG signal recorded on the monitored brain and a normal reference EEG
- [12] that distinguishes between ictal and seizure-free EEGs using empirical mode decomposition and Fourier-Bessel expansion
- [13] that fuses features extracted from the EEG and its accompanying ECG to detect temporal lobe epileptic seizures
- [14] that uses the fractal dimension to distinguish between normal EEGs and EEGs of dementia patients

However, these methods are usually tested on different, small sets of data therefore their results remain hard to reproduce, assess and interpret with any

certainty.
There is little doubt, based on this, that some support needs to be given clinicians to make the diagnosis process faster and more accurate. Providing a medical data sharing platform is one of the possible solutions to improve the diagnosis process. A shared medical repository would also help researchers in their task of developing more accurate and reproducible (semi)-automated medical data interpretation methods in that it would provide a large trove of "standard" data.

1.2 Benefits of a shared medical data repository

So how would building a medical data sharing platform help improve the diagnosis process? Assuming solid mechanisms are put in place to allay privacy concerns (due to the sensitive nature of the data), sharing the data would:

- make the patient records and test results available to all physicians and specialists treating the patient
- ease the access to his/her medical history for all the different treating physicians and specialists
- improve patient data security by managing it in one (possibly distributed) repository whose security can be more easily maintained and protected better against attacks than data stored in islands of data
- facilitate the automated analysis of medical data through machine learning algorithms
- benefit research as it would facilitate the construction and reuse of datasets thus improving the comparability and reproducibility of the results

Sharing such complex medical data and making it easily available to clinicians may also promote the collaboration between clinicians and make them reach collegial thus more accurate data interpretations and diagnoses.
Making the same data available to all physicians and specialists involved in a patient's care is especially crucial and would improve the diagnosis and care process. Having the physicians all know which tests have been performed and which potential diagnoses have been reached and discarded would guide them in their diagnosis and choice of course of treatment, make them explore

previously unexplored diagnoses if need be and come up with accurate diagnoses and courses of treatment faster while avoiding pitfalls such as unnecessary medical tests, potential misdiagnoses and overlooked diagnoses. The same benefits may be obtained if clinicians have access to the patient's history: sharing the data would make that history accessible. Ultimately, medical data sharing would improve patients' outcomes and quality of life.
A shared medical database would be a trove of data on which competing machine learning algorithms and analysis techniques could be tested and easily compared, interpreted and reproduced. Furthermore, most machine learning techniques benefit from being tested on big datasets. A shared medical data store makes that data available for analysis and researchers can evaluate automated diagnosis methods as well as the benefit-risk balance of particular treatments or diagnosis tests, in keeping with the principles of evidence-based medicine. Medical tests data is very often interpreted visually by trained specialists. Such visual interpretation is for instance the golden standard in EEG interpretation. But not only is such a visual interpretation expensive, tedious and time-consuming, it is also error-prone. This is partly due to the quantity of data to interpret. For instance, the interpretation of each routine 20 minute EEG requires the perusal of 109 A4-pages, following the guidelines of the American Clinical Neurophysiology Society [15], keeping in mind that while most EEGs are routine ones, many are longer than 20 minutes and up to days of recording (eg ICU patients' EEG monitoring) and that each EEG recorded -the scale of which is visible in Table 1.1- has to be interpreted within days of its recording. But it is also due to data specificities. EEG recordings, for example, are rife with non-specific patterns, artifacts as well as age or context-dependent patterns. For example, a chewing or toothbrush artifact may be mistaken for an epileptic seizure or the presence of delta waves-i.e waves with a frequency of 3HZ or less- may be found normal in infants, children and deeply asleep adults[3] or pathological in awake adults[4].
Additionally, the data and patient privacy would be more thoroughly secured and protected by storing medical data in a single repository and then sharing it. Last but not least, storing medical data in a single data warehouse to which authorized users are given access also minimizes the risk of system failure and parts of data becoming totally unavailable.
Analyzing the US healthcare system, the MGI report cited earlier ([8]) concludes that collecting, sharing and analyzing medical data (big data) offers

[3] younger than 65
[4] younger than 65

Table 1.1: Some statistics on medical data (Source: OECD Report 2011 ([16]), figures from 2009, last year for which records are available)

	Netherlands	**USA**	**OECD**[3]
EEG[1]	100,000 167GB	N/A	N/A
MRI[2]	726,000 15.9TB	28 million 614TB	42 million 921TB
CT[2]	1.1 million 36.7TB	70 million 2.3PB	104.5 million 3.4PB

huge premiums. Such premiums include drastically reducing health care costs and waste and improving patient outcomes and quality of life through easing the deployment of clinical decision systems, facilitating comparative effectiveness studies, increasing data transparency and even allowing remote patient monitoring.

1.3 Building a shared medical data repository: challenges

The MGI report cited earlier ([8]) also points out significant technical hurdles to overcome, on top of legal hurdles, before medical data can be shared and analyzed properly and its full potential uncovered. Among those technical hurdles, standardizing data formats, guaranteeing systems' interoperability, integrating already existing, fragmented and possibly heterogenous datasets and providing sufficient storage are cited.
The scale of the data that is generated and has to be interpreted in the healthcare system is indeed huge as highlighted in Table 1.1. Furthermore, the medical data we seek to share through a repository is a collection of very diverse sets of data:

[1] Assuming standard 20-minute EEGs stored in EDF+ format. Average size per file 13.7MB.

[2] Assuming average size of 23MB per MRI and 35MB per CT

[3] Based on data from OECD countries for which data is available for exams performed in and outside of hospitals i.e the USA, Greece, France, Belgium, Turkey, Iceland, Luxembourg, the Netherlands, Canada, Denmark, Estonia, the Czech Republic, the Slovak Republic, Chile, Israel and South Korea

- textual data describing patient symptoms, patient course of treatment, doctor observations or recommendations
- raw test data mainly consisting of sensor data (eg CT scans, MRI scans, EEG recordings)
- test data interpretation done by a specialist or with the help of semi-automated interpretation methods

Such data is also collected and stored at various locations (or islands of data) as clinicians order diverse batteries of tests, often repeatedly performed to confirm a finding or test new hypotheses.
As we mentioned earlier, the huge amounts of medical data to be interpreted are generally stored at various locations. Currently, medical data is scattered across different hospitals, clinics, private practices and diverse research institutes or universities, with data often being passed from one person to another physically on hard drives or other external storage devices. As a result, the risk of data being exposed to unauthorized people as well as the likelihood of inconsistent copies of the same data being created are high. The data is harder to trace and it is not straightforward to determine what kind of data is available, where it is available and to who it has been made available.
Once the data is shared through a suitable platform, one has to be able to access the data in response to queries. The following queries are examples of possible queries on EEG and MRI data:

1. find EEGs of patients aged between 20 and 30 and showing patterns consistent with temporal lobe epilepsy
2. find EEGs showing rhythms associated with consumption of barbiturates
3. find sequences of EEGs where the mu rhythm appears
4. remove artifacts from sequence of interest Y
5. show an EEG with similar patterns to that of patient X
6. show the tumor area in the MRIs of patient X after the start of treatment Y

Obtaining a simple answer to this set of queries would require the data to be heavily and precisely annotated and tagged. But what if the annotations are

scarce or not available at all? Besides, the whole process of manually annotating and tagging each and every part of the medical tests datasets is time-consuming and error-prone. Feature extraction techniques need to be used to respond to all these queries as they can process the raw data so as to:

- define a set of clinical features representative of a particular pathology (eg epileptic features present in channels corresponding to the temporal lobe of the brain in query 1, consumption of barbiturates in query 2)
- analyze the EEG in terms of frequencies, retrieve sequences showing the presence of some kind of cerebral wave (the mu rhythm in query 3)
- remove artifacts from sequences based on features defining artifacts (query 4)
- help establish a diagnosis by comparison (in query 5, a similarity measure between EEGs needs to be defined)
- segment the brain into chemically-distinct structures (healthy tissue and tumorous tissue in query 6)

Moreover, as previously stated machine learning algorithms may be used to perform (semi)-automated data interpretation. So whether it be in response to queries or in order to perform (semi)-automated interpretation of the data, the data shared through a medical repository needs to be easily accessible for further processing, ideally on the sharing platform itself. This poses two additional challenges. The medical data processing methods are usually computationally expensive. For example, computing the matrix inner product $\mathbf{AA^T}$ (with $\mathbf{A} \in \mathbb{R}^{n \times D}$), which is a mainstay of many similarity measures and distance-based clustering methods as well as feature reduction methods such as principal component analysis, has a complexity of $\mathcal{O}(n^2 D)$. This means that, if you do not reduce the EEG dimensionality by extracting features, computing such an inner product for a standard 20-minute EEG following the 10/20 system (therefore comprising 19 data channels) and sampled at 250Hz would require $(20 * 60 * 250)^2.19 = 1.71 \times 10^{12}$ operations. This computation is likely to take a while. Another example is that of the Fourier transform, which is frequently used as a first analysis step for EEGs. The most used Fourier transform computation algorithm is known as the Fast Fourier Transform (FFT) and has a complexity of $\mathcal{O}(n \log(n))$. Therefore, applying the FFT algorithm to a single 20-minutes standard EEG without dimensionality reduction requires

$5,700,000 \log(5,700,000) \approx 38,508,487$ operations to be performed. The number of operations required to compute the FFT or inner product on an EEG would obviously be reduced if features are extracted from the EEG to reduce its dimensionality but the question would shift to determining the set of relevant EEG features for the tasks at hand, which is far from straightforward and highly dependent on the application. Both the inner product and FFT examples show that, without carefully considering how to make the processing of the data available as efficiently as possible, applying even simple feature extraction, clustering or other machine learning methods quickly becomes unmanageable as the amount of data available grows. The second challenge is that these methods inevitably add to the uncertainty of the interpretation, though the added uncertainty would, in this case, be quantified unlike the uncertainty arising from the visual interpretation or from the raw (sensor) data itself.

Small summary
When dealing with medical data, we have to deal with data that is:

- scattered and hard to trace (islands of data problem)
- very diverse
- extremely large (see Table 1.1)
- hard to interpret
- difficult to process efficiently and within reasonable times with machine learning techniques
- highly uncertain

So the question is- and this is our research question: *how can we build an integrated sharing and processing platform for medical data to support the medical diagnosis process?*

1.4 Research question

As outlined earlier, our research question is how to build an integrated data sharing and processing platform for medical data to support and ease the medical diagnosis process. This research question can be split into three parts.

The first part is the building of the data sharing platform. This platform has to be able to deal with huge amounts of data. Ideally, medical data should be shared globally but this is unlikely to happen in the foreseeable future in particular for legal reasons. However, medical data should at least be shared on a national level. The annual recorded medical data (EEG, MRI and CT data) on a national scale ranges from a dozens of terabytes (eg the Netherlands) to petabytes of data (eg the USA) as shown in Table 1.1. So the scale of data the storage platform to be built has to deal with is national scale amount of data i.e up to petabytes of data. Furthermore, since medical data is highly heterogeneous, the storage platform has to be able to store diverse and possibly unstructured types of data eg (multidimensional) time series such EEG, ECG or MEG data or images such as MRI, CT and PET scans
Finally, medical data has to be accessible for further processing using for instance machine learning techniques. So what type of platform would fit these requirements?

The second part of the research question concerns data retrieval. As outlined earlier, the queries that need to be served by the shared medical data repository are semi-structured queries that need features to be extracted from the raw data to be answered. So the second part of the research question would be: what kind of feature extraction techniques can be used to index the data so that the data is easily retrieved and accessed in response to semi-structured queries?

Whether it be to index data for easy retrieval or to interpret data to help with the diagnosis process, feature extraction and machine learning techniques will need to be used. Such techniques add uncertainty on top of the uncertainty already existing in raw medical test data. This uncertainty in particular affects the labeling of data e.g the labeling of EEG events or the labeling of an EEG with a possible diagnosis used as part of reaching a conclusion and final diagnosis. So one has to be able to quantify such uncertainty since it affects the decision-making process (diagnosis process and patient treatment and care process). And one also has to allow the addition of new evidence such as user feedback to quantify and refine the uncertainty estimates. So the third part of the research question is: how do we combine diverse sources of evidence- one of which is user feedback- to quantify and possibly reduce the uncertainty linked to discrete variables such as a medical diagnosis?

Our initial research question therefore went from how to build an integrated data sharing and processing platform for medical data to support and ease the medical diagnosis process to the following three subquestions:

- how do we design a data sharing platform so as to fit the previously highlighted constraints?
- what types of machine techniques should we use use for data indexing and retrieval in response to semi-structured queries?
- how do we combine evidence such as user feedback to quantify and possibly lower the uncertainty attached to discrete variables-used in the decision-making process- such as the medical diagnosis variable?

1.5 Contributions

There are four main contributions in this thesis. First, we show that a possible storage framework for medical data would include two parts communicating with each other:

1. a Hadoop cluster where raw data files stripped of patient information for confidentiality and security would be stored and processed
2. a query and search layer where metadata such as patient information, information obtained from feature extractors, indexes, lineage and versioning information and uncertainty information could be stored. Such metadata could be stored in an RDBMS or an XML database (for more flexibility in the storage format)

And we provide a proof of concept for the suitability of a Hadoop cluster as a storage and processing platform for raw medical test data files.
As a second contribution, we propose a method relying on a fractal-dimension-based similarity measure that could be used to retrieve EEGs once stored in the medical data sharing platform. While the features chosen try to achieve good retrieval rates for EEG data, the features used are generic time-series properties that could be investigated for the retrieval of other medical time series data. Furthermore, the principle of the EEG-similarity search approach can be applied for other types of data in particular other medical time series.

Because the medical diagnosis process is incremental, uncertain and evidence-based (eg evidence obtained through user feedback or (semi)-automated medical data interpretation methods), our third contribution is a Dempster-Shafer theory-based model that quantifies of the uncertainty attached with medical diagnoses using incrementally obtained user feedback and other sources of evidence (eg input from (semi-)automated diagnosis techniques). The model built takes into account the fact that all sources of evidence are not necessarily equally reliable and that the variables receiving new feedback/evidence may be derived or linked to other variables through lineage.
Finally, we show how the built Hadoop-based storage platform, the explored data indexing and retrieval methods and the evidence combination model fit together and can be used in semi-structured queries processing.

1.6 Thesis organization

The whole thesis takes EEG data as example of medical data and focuses only on this type of data so we introduce some background information on EEG data in Chapter 2.
We then study the suitability of Hadoop as a medical data sharing and processing platform in Chapter 3. And, in Chapter 4, we build Dempster-Shafer theory-based model that quantifies the uncertainty attached with medical diagnoses using incrementally obtained user feedback and other evidence (eg input from (semi-)automated diagnosis techniques). The model built takes into account the fact that all evidence is not necessarily equally reliable and that the variables receiving new feedback/evidence may be derived or linked to other variables through lineage.
Chapter 5 deals with similarity search in EEG data. It investigates two feature extraction methods that may be used to index EEGs so as to allow their fast retrieval in response to common user requests:

- fractal interpolation and fractal dimension computation for EEG compression and classification
- event detection in EEG followed by rule-based classification of EEGs (eg. a normal EEG contains no events or events classified as artifacts)

We also show, in chapter 5, how the built Hadoop-based storage platform, the explored data indexing and retrieval methods and the evidence combination model fit together and can be used in similarity search.

Chapter 6 concludes the thesis and suggests possible avenues for future research in the domain of medical data storage and processing.

Acknowledgments

The EEGs used in most of the chapters of this thesis (Chapters 3 and 5) were kindly provided by Prof. Dr. Ir. Michel van Putten (Dept. of Neurology and Clinical Neurophysiology, Medisch Spectrum Twente and MIRA, University of Twente, Enschede, The Netherlands), who we also thank for useful insights on EEG data.

CHAPTER 2

Background on EEG Data

The purpose of this thesis is to build a storage and processing platform for medical data, with EEG data chosen as an example of complex medical data to accomplish such a task. This chapter provides some background on EEG data.

2.1 General principles

During an EEG recording, the brain electrical activity is captured through several electrodes (21 in the 10-20 system) placed on the scalp. The signal is then amplified (the amplitude of the resulting signal is usually 10^6 times that of the original signal, the signal amplitudes being of the order of the μV), filtered and if recording the data onto a computer discretized (at a certain sampling rate, usually around 250 Hz).
The skin-electrode impedance has to be monitored closely since an impedance exceeding 5 $k\Omega$ results in artifacts in the EEG recordings. Therefore, prior to electrodes' placement, the points of contact skin-electrodes are scrubbed to make sure that the skin-electrode impedance does not exceed 5 $k\Omega$ during the measurement and to remove dead skin cells and dirt.
The electrodes are placed on the skull according to a standard known as the International 10/20 System. The 10/20 System relies on the calculation of distances between fixed points on the head: the electrodes are placed at points that are 10% and 20 % of these distances (see Figure 2.1). Once the electrodes have been positioned correctly, they can be connected in different ways/montages according to, for instance, the underlying pathology or brain zone explored. For example, a derivation with small distances between electrodes (the distance between two given electrodes does not exceed 3 cm(see Figure 2.1)) can be used when trying to scan a narrow zone of the brain and conversely a

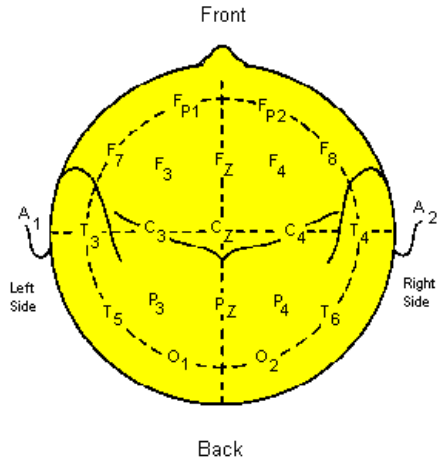


Figure 2.1: Positioning of the EEG electrodes according to the International 10/20 System (Source: http://faculty.washington.edu/chudler/1020.html)

derivation with big distances between electrodes (i.e exceeding 6 cm) might be used when trying to detect the brain's basal activity. The American EEG Society ([17]) suggests the use of three montages as standard for clinical practice. The first two montages are bipolar montages,i.e montages connecting pairs of active adjacent electrodes and computing the differences of potential between them. The first bipolar montage to be used is called the bipolar longitudinal montage. In this montage, the brain is scanned from the front to the back with the right and left sides of the brain being explored simultaneously, which means that Fp1 is connected to F3, Fp1 to F7, F3 to C3, etc. The second bipolar montage suggested is the transverse montage. In this montage, starting from the electrodes F7, T3 and T5, the brain is explored from left to right and from front to back (the electrodes Fp1, Fp2, O1 and O2 are not used in this derivation). The third suggested montage is called the referential montage. The differences in electric potential are measured between an active electrode and an electrode of reference (for example electrodes A1 and A2). When this derivation is used, the brain is explored from front to back and/or from left to right by connecting each of the active electrodes to the electrode of reference. Figures 2.2 and 2.3 shows a few examples of EEGs with referential derivation. See [18, 19, 20, 21] for more details.

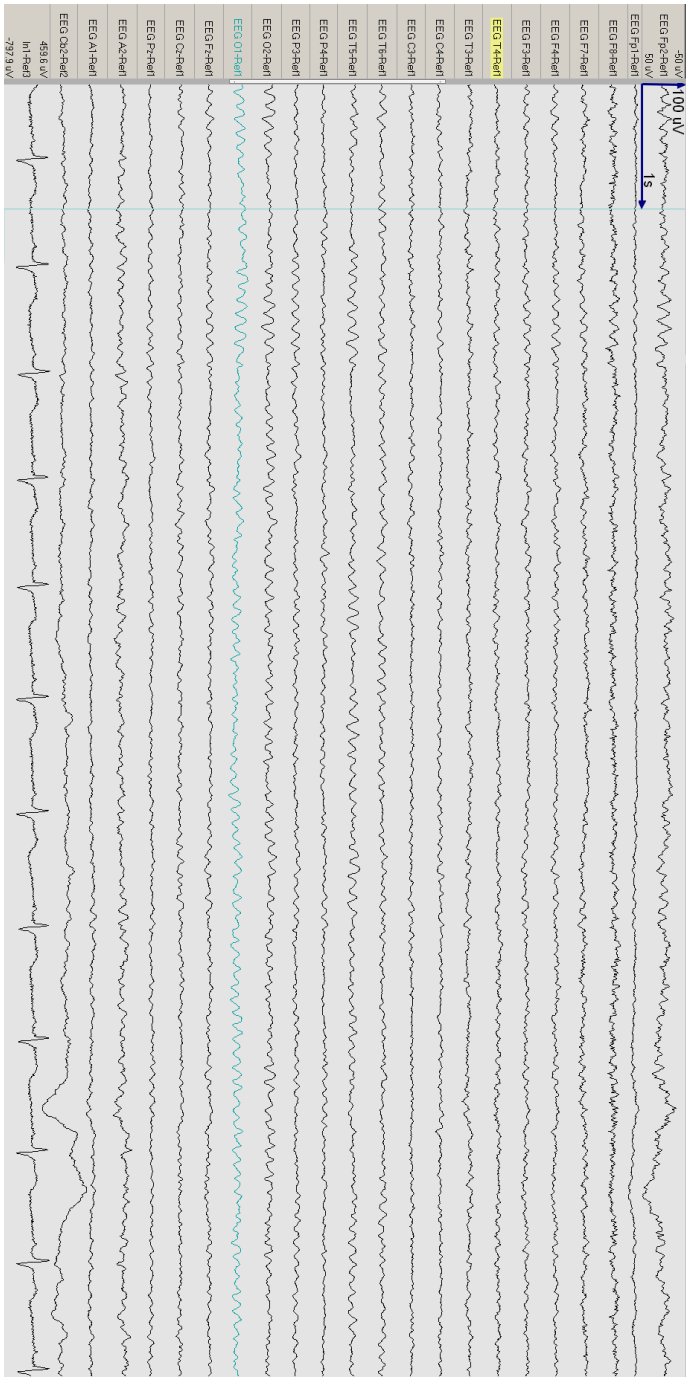


Figure 2.2: Normal eyes closed EEG segment (10 seconds of recording, referential montage, adult patient)

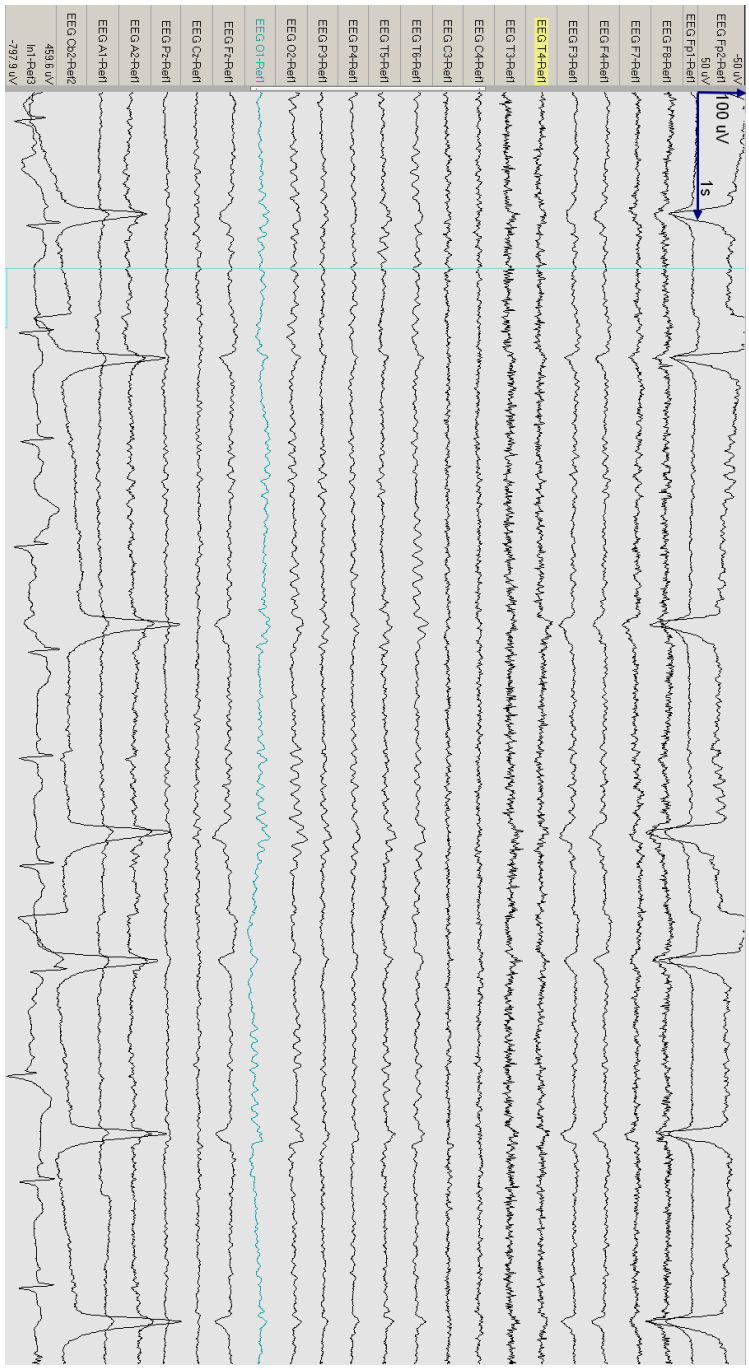


Figure 2.3: Normal eyes open EEG segment (10 seconds of recording, referential montage, adult patient)

A standard EEG recording includes several sequences:

- a sequence of about 15 minutes in which the patient, at rest, opens or closes his/her eyes according to the technician's instructions.
- a sequence of brief stimuli (visual, auditive and nociceptive) followed by periods of rest
- an activation sequence in which the patient undergoes hyperpnea (hyperventilation) or stimulation by stroboscope (also called photic stimulation)
- a rest sequence

Hyperventilation and photic stimulation are methods used to accentuate or provoke EEG abnormalities, which may not be otherwise visible. If the standard EEG recording does not show any abnormality but the clinical findings suggest otherwise, other EEGs may be recorded under sleep or after a 24-hour sleep deprivation as abnormalities are more likely to crop up in these types of recordings.

2.1.1 Cerebral waves

An EEG recording (in a channel) can be classified into several types of cerebral waves characterized by their frequencies, amplitudes, morphology, stability, topography and reactivity. The waves are classified in bands of frequency, in particular:

- δ wave band for waves whose frequency is lower than 3.5 Hz
- θ wave band for for waves whose frequency is between 4 to 7.5 Hz
- α wave band for waves whose frequency is between 8 to 13 Hz
- β wave band for waves whose frequency is between 13 to 30 Hz
- γ wave band for waves whose frequency is higher than 30Hz

The α wave band consists in rhythmic waves of amplitude comprised between 20 and 100 μV distributed in a bilateral and synchronous fashion in the posterior regions of the brain. The amplitude of those waves is maximal when the eyes are closed. The α activity is blocked when the eyes are opened or when

something that requires attention is being performed. The α activity can disappear during a complex mental activity and be replaced by fast β activity. In contrast, the basal α activity is reinforced during the few seconds after closing the eyes. The α activity is normal when present in awake adults.
The α activity can hardly be identified in normal subjects in 10% of the cases. In this case, the posterior basal activity is then replaced by a slow activity of amplitude between 10 to 30 μV.
Theta waves appear as a result of drowsiness. They also appear frequently in infancy and childhood with their amount decreasing as the brain matures. Therefore the EEG of a normal awake adult contains very weak θ activity. An excess of θ activity (diffuse or localized) in awake adults is considered abnormal.
Just as theta waves, delta waves are a marker of brain maturation and, as such, appear frequently in children EEGs. Delta activity decreases with brain maturation. As a result, the EEG of normal awake adults contains almost no δ activity. The only strong δ activity that appears in adulthood is during sleep. The presence of delta activity in an awake adult, be it in a diffuse or localized manner, always signals an underlying pathology.
A β activity of frequency comprised between 13 to 30 Hz and amplitude lower than 20 μV can occur often asynchronously in the middle regions of the two brain hemispheres. β rhythms can also be observed when certain medications (eg barbiturates, benzodiazepines) are used.

2.1.2 Artifacts

As any recorded signal, an EEG can be marred by several artifacts due to:

- physiological phenomena such as muscle activity (eg jaw muscles clenching or tremor), eyes or head movements during recording, skin-electrode impedance exceeding 5 $k\Omega$, perspiration or hyperventilation accompanied by body movements

- the equipment such as problems with electrodes, moving connection wires or faulty connection

In some cases, the morphology of these artifacts can be similar to some pathological wave patterns (e.g epileptiform transients (ET)). For instance, the so-called ECG (electrocardiogram) artifact produces patterns that may be misinterpreted as sharp waves or spike discharges in particular if the ECG rhythm is

irregular. Therefore an ECG is usually recorded along the EEG so as to be able to detect any contamination of the EEG signal by the ECG signal and avoid the erroneous conclusion of the presence of epileptiform abnormalities. The ECG artifact occurs when ECG potentials (that measure cardiac activity) have a large enough amplitude to be detected by cerebral electrodes. For more details on EEG recordings and EEG patterns, see [22, 23, 18, 24, 19, 20, 25, 21].

2.2 Applications

EEG recordings are useful tools for the diagnosis of several neurological disorders and abnormalities and they are, in particular, used to :

- detect epileptiform patterns
- caracterize the type of epilepsy once a diagnosis of epilepsy has been reached based on the patterns found in the EEG
- localize the origin of seizures
- indicate the most appropriate (epilepsy) medication to prescribe
- check the effect of epilepsy medication
- control anesthesia depth during surgeries
- locate brain areas damaged by a stroke, tumour or head injury (though it's been mostly replaced by CT and MRI scans for this application)
- monitor cognitive engagement, alertness, coma or brain death
- investigate sleep disorders
- monitor the brain development
- investigate mental disorders

2.3 EEG file format

Several file formats are used to store EEG scans: formats such as Neuroscan EEG files (*.eeg, *.cnt, *.avg) or Biosemi BDF files (*.bdf). An increasingly popular file format is the EDF+ file format. EDF+ files were designed to store and

HEADER

8 ascii	version of this data format(0)
80 ascii	local patient identification
80 ascii	local recording identification
8 ascii	startdate of recording (dd.m.yy)
8 ascii	starttime of recording (hh mm ss)
8 ascii	number of bytes in header record
44 ascii	reserved (+C for continuous signals, +D for discontinuous signals)
8 ascii	number of data records (-1 if unknown)
8 ascii	duration of a data record, in seconds
4 ascii	number of signals (ns) in data record
ns * 16 ascii	ns * label (e.g. EEG Fpz-Cz or ECG)
ns * 80 ascii	ns * transducer type (e.g. AgAgCl electrode)
ns * 8 ascii	ns * physical dimension (e.g. uV)
ns * 8 ascii	ns * physical minimum (e.g. -500)
ns * 8 ascii	ns * physical maximum (e.g. 500)
ns * 8 ascii	ns * digital minimum (e.g. -2048)
ns * 8 ascii	ns * digital maximum (e.g. 2047)
ns * 80 ascii	ns * prefiltering (e.g. HP:0.1Hz LP:75Hz)
ns * 8 ascii	ns * nr of samples in each data record
ns * 32 ascii	ns * reserved

DATA RECORD

number of samples[1]*sample value(2-byte integer) first signal of the record
number of samples[2]*sample value(2-byte integer) second signal of the record
.
.
.
number of samples[ns]*sample value(2-byte integer) last signal of the record

Figure 2.4: Structure of an EDF+ file

exchange digital recordings such as EEGs, EMGs, Evoked Potential studies. Each EDF+ file contains two parts: a header record followed by a collection of data records. The header of an EDF+ file typically contains information about the patient as well as the technical characteristics of the EEG signal, such as the type of recording, the type of recording equipment used, the number of epochs (called data records) contained in the transcribed EEG, the duration of the EEG epoch, the EEG channels' labellings, the number of data points per epoch and the number of signals in the EEG. The data records contain consecutive fixed-duration epochs of the EEG recording. In other words, the second part of the file is a succession of data records representing time slots that each contain the EEG signal values for all EEG channels for that particular time slot. Annotations on an EEG signal can be stored in an EDF+ file as an additional signal. The structure of an EDF+ file is depicted in figure 2.4. For more details on EDF (ie the file format which EDF+ extends) and EDF+ file formats specifications,

see ([26, 27]).

2.4 EEG automated interpretation

The visual inspection of an EEG recording by a neurologist is the current gold standard of EEG interpretation. This not only requires skills but is also time-consuming, especially since there is a trend towards recording lengthy EEGs as it has been shown that the detection rate of epilepsy improves with the length of recording ([28, 29]). Furthermore, [30] shows that EEG interpretation varies widely between experts: eight experts EEG interpreters were asked to mark epileptiform discharges in twelve short EEG recordings but 38% of the discharges were marked by only one expert and only 18% by all experts.
Such considerations have led to the development of several automated EEG classification and interpretation methods.
Some focus on discriminating EEGs between normal EEGs and EEGs of a particular condition: dementia in [14], epilepsy in [12, 13]. Others try to detect specific patterns in EEGs such as epileptiform discharges in [31, 32, 33, 34], seizure activity ([35]), EEG background activity in [36] or sleep stages in [37, 38].
Quantitative EEG analysis is also being used to obtain prognosis information for patients with ischaemic stroke([39, 40]).
In other approaches ([41]), "relevant" EEG features are selected, quantified and visualized through time to be presented to a practitioner who then interprets them and their variations to derive conclusions on the EEG.
For a more thorough review of EEG automated interpretation methods, in particular the detection of epileptiform discharges, see [42].

CHAPTER 3

Medical data sharing and processing with Hadoop

The contents of this chapter have been published in the proceedings of the 2014 International Conference on Brain Informatics and Health (BIH 2014) ([43]).

3.1 Motivation

We showed in the introduction (Section 1) that there were big benefits to sharing medical data in a medical repository not least improving patients' outcomes and quality of life, reducing healthcare waste and costs and tightening patient data security. We also showed that due to the amount of medical data and its complexity it would be helpful to automate at least part of the diagnosis process so as to ease the clinicians' workload and improve their performance. And we pointed out the constraints any potential design for a medical repository need to take into account: the distributed nature of medical data, its heterogeneity and size, the diversity of file formats and platforms used across healthcare institutions and data accessibility for further complex processing. So the question is now to find/design a suitable platform that fits these constraints. This chapter seeks to demonstrate, using EEG data as example of medical data, that a rather low cost technical solution (and possible storage platform for medical data) that fits the required constraints and requires minimal changes to current state of the art storage and processing techniques already exists: the Hadoop platform.

3.2 Contributions

This chapter gives a proof of concept for an EEG repository by :

- explaining why Hadoop fits the constraints imposed on potential medical data repositories
- showing how to store EEG data in a Hadoop framework
- proving that EEG data can be analyzed on national scale on Hadoop by designing and benchmarking a representative machine-learning algorithm

3.3 Related work

Hadoop has been found a viable solution for storing and processing big data similar to medical data, such as images in astronomy ([44]) or power grid time series, which unlike medical time series, are unidimensional time series ([45]). [46] is, to the best of our knowledge, the first paper to consider storing medical data and EEGs in particular with Hadoop and show it is a promising solution in need of more testing. [46] suggest exploring the "design and benchmarking of machine learning algorithms on [the Hadoop] infrastructure and pattern matching from large scale EEG data". This is one of the goals of this chapter.

3.4 Hadoop: a good fit for medical repositories' constraints

3.4.1 Introduction to Hadoop

Hadoop, an open source platform managed by the Apache open source community, has 2 core components: the Hadoop Distributed File System (HDFS) and the job management framework or MapReduce framework. The HDFS is designed to reliably store huge files on all cluster machines. Each HDFS file is cut into blocks and each block then replicated and stored at different physical locations in the cluster to ensure fault tolerance.
The HDFS has a master/slave architecture with one master server called *Namenode* managing the filesystem namespace and regulating the file access by clients and multiple slave servers (one per cluster node) called *Datanodes* managing the storage in the nodes they run on. The *Namenode* maps the file blocks

to the *Datanodes* and gives the *Datanodes* instructions to perform operations on blocks and serve filesystem clients' read and write requests.
The Hadoop MapReduce framework also has a master/slave architecture with a single master called *jobtracker* and several slave servers (one per cluster node) called *tasktrackers*. MapReduce jobs are submitted to the *jobtracker*, which puts the jobs in a queue and executes them on first come/first serve basis. The *jobtracker* assigns tasks to the *tasktrackers* with instructions on how to execute them.

3.4.2 Hadoop and parallel data processing: the MapReduce model

MapReduce is a programming model for data-intensive parallelizable processing tasks (introduced in [47]) designed to process large volumes of data in parallel, with the workload split between large numbers of low level commodity machines. The MapReduce framework, unlike parallel databases, hides the complex and messy details of load balancing, data distribution, parallelization and fault-tolerance from the user in a library, thus making it simpler to use the resources of a large distributed system to process big datasets. The MapReduce model relies on 2 successive functions to transform lists of input data elements into lists of output data elements: a *mapper* function and a *reducer* function. Each input data element is transformed into a new output data element by the *mapper*. The transformed elements are then aggregated by the *reducer* to return a single output value. A simple example is files word count: in this case, the *mapper* associates a number of words to each of the input files while the *reducer* function sums the values obtained during the mapping step.

3.4.3 Hadoop for medical data storage

The Hadoop platform provides a solution to the technical hurdles outlined by the MGI report ([8]) described earlier (Section 4.1).
First of all, Hadoop was designed to scale with large data. It is currently being used at Facebook to store about 100PB of user data, i.e data much bigger than national scale medical data which ranges from dozens of terabytes (eg the Netherlands) to petabytes of data (eg the USA) annually as shown in Table 1.1. So Hadoop can easily handle national scale amount of medical data.
Moreover, Hadoop can store heterogeneous formats of data, in particular unstructured data, and if there is a method to extract the data from the files that

store it [1], the data can then be fed to Hadoop MapReduce for further analysis and processing.
Hadoop is also tolerant to node failure. The HDFS relies on replication (by default 3 copies on 3 Datanodes per file block) to ensure file blocks are not lost if a data server fails. If a Datanode fails and some data blocks have less than a set minimum of copies, the Namenode orders the replication of the affected blocks in some available Datanodes to bring back the replication factor of the blocks to safer levels. The probability of losing a block in a 4000 nodes' cluster in a day (respectively in a year) in the case of uncorrelated failures of multiple nodes is about 5.7×10^{-7} (respectively 2.1×10^{-4}) ([48]). At Yahoo! in 2009 for example, only 641 blocks were lost out of 329 million on 17720 nodes i.e a loss rate of 1.9×10^{-4}% ([48]). The only problem left is the Namenode as the HDFS is unusable if the Namenode fails. Namenode crashes rarely occur though ([49])(1 in 4 years at Facebook) and solutions limiting the crash impact are already being deployed. One such solution is the AvatarNodes in use at Facebook: 2 AvatarNodes, an active and standby one, replace the unique Namenode and receive the Datanodes messages in its stead. The Standby AvatarNode thus contains up-to-date information about block locations and can be started in under a minute to replace the Namenode (or Active AvatarNode) if it fails. This solution cuts cluster planned downtime by 50%. Data stored with Hadoop will therefore be constantly available.
Hadoop was built for parallel processing (via MapReduce described in Section 3.4.2) and we study the feasibility EEG data processing with Hadoop with the example of feature selection by exhaustive search in Section 3.5.

3.4.4 Hadoop and EEG storage

An EEG is a multidimensional time series obtained by capturing the brain's electric activity with scalp electrodes. Figure 2.2 (in Chapter 2) shows an example of EEG. The increasingly popular EDF+ format is used to store EEGs and contains all the information about the EEG recording, both metadata in a header encoded in UTF-8 and raw data in binary format. The metadata includes patient information and EEG signal technical attributes (eg equipment details and sampling rate). Annotations on the EEG, such as context of recording or EEG events labels, may also be stored in the EDF+ file. See [27] for

[1] Such methods currently exist at the sites where the different types of data are stored. There is,at most, a need to translate those methods into Java, Python, Perl or any other language that can be interfaced with Hadoop.

format details.
HDFS does not call for any set file format, so we store EEGs in EDF+ in HDFS. We anonymize EEGs before storage for security reasons. Keeping EEGs as EDF+ files has many advantages. No additional data formatting is needed and existing tools for EDF+ files, eg. visualization tools, can still be used. And as EDF+ files are mainly binary files, the size of the stored EEGs is small: 2500 EDF+ files (dataset 1 in Section 3.6 and Table 3.1(a)) i.e to about 2 years of EEG data at the local hospital take up 46.5GB whereas the same data [2] would take up 1TB when in a relational database.

3.5 EEG feature selection with Hadoop

EEG interpretation is arduous even for trained specialists due to the mass of data to interpret [3] and non-specific, age or context-dependent patterns and artifacts. For example, in the absence of observation, the EEG patterns recorded on a patient brushing his teeth can be mistaken for epileptic seizure activity ([50],p.112). Machine learning-based methods ([14, 51]) are being developed to ease the interpretation for clinicians, though the methods' scalability remains an issue. Instead of reducing algorithm complexity as in most studies aiming to lower the computational cost of machine-learning methods, we opt for using more commodity hardware with Hadoop and show, here, with EEGs as example, that parallelizable machine learning tasks, in particular tasks that may be translated into a sequence of *map/reduce*, can be run in manageable times.

3.5.1 Feature selection as example EEG machine learning algorithm

Most automated EEG data interpretation methods classify or cluster EEGs and select suitable features for classification/clustering (eg. fractal dimension in [14, 52]) prior to it. Other approaches ([41]) select, quantify, visualize some "relevant" EEG features through time and present them to a practitioner who then interprets them and their variations to derive conclusions on the EEG. So the key task in the automated interpretation of EEG is feature selection so we

[2] with one table for metadata, one table for raw data and one tuple per raw data point
[3] a routine 20 minute EEG fits in 109 A4-pages with the guidelines of the American Clinical Neurophysiology Society [15]

pick a feature selection algorithm on EEG as example of machine-learning algorithm to determine whether Hadoop is suitable for medical data processing compared to other more traditional frameworks. We purposely choose an algorithm with exponential complexity for feature selection (exhaustive search) as achieving manageable execution times with Hadoop for this worst-case algorithm would entail achieving even more reasonable execution times for more common less computationally expensive algorithms. The goal of this study is not to evaluate the accuracy of the feature selection algorithm but to test whether running feature selection (as a sample machine-learning algorithm) on Hadoop has any benefits compared to using more traditional processing platforms.

3.5.2 Tested features and rationale for the choice of features

To test the feature selection algorithm, we choose a mix of 9 clinically-relevant and more general time-series features shown to be relevant for EEG processing in literature:

- 4 features computed in the time domain (fractal dimension, mean amplitude, amplitude standard deviation, normalized Hjorth mobility and complexity[4])
- and 5 in the frequency domain (frequency bands percentages (α band,β band,θ band,δ band)[5], the α to δ ratio, high to low frequency ratio (high frequency being frequencies above 25Hz), brain symmetry index (BSI) and spectral entropy)

These features detect many pathologies and patterns: EEG asymmetries as in focal seizures or hemispheric ischemia with the BSI defined in [53], temporal lobe seizure with the Hjorth mobility and complexity([51]), high frequency artifacts with the high to low frequency ratio ([54]), hypofunctional patterns with the α to δ ratio and iso-electric ([54]), low-voltage EEGs with the mean amplitude ([54]). The fractal dimension separates normal sequences and other sequence types ([52]) and normal EEGs and Alzheimer patients EEGs ([14]). An extra feature, the nearest neighbour synchronization (mNNC) (defined in

[4] 2-dimensional feature

[5] The EEG waves are grouped by frequency in 4 main bands: δ band for frequencies from 0.5 to 4 Hz, θ band for frequencies from 4 to 7 Hz, α band for frequencies from 7 to 12 Hz and β band for frequencies from 12 to 30 Hz. The frequency band percentage is therefore a 4-dimensional feature.

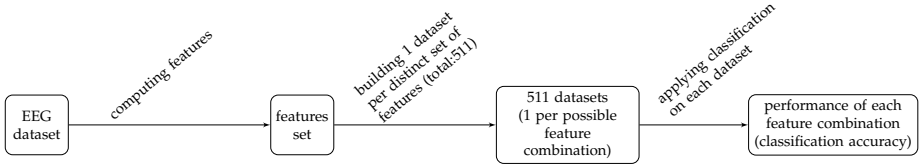


Figure 3.1: EEG feature selection steps

[41]), used to detect seizures ([54]), sleep or encephalopathies ([41]) [6] is computed in the feature computation step (to measure scalability) but not used for classification.
Each of the 9 features can be picked alone or in combination with a variable number of the other features. So there are $\sum_{i=1}^{9} C_9^i = 511$ distinct possible ways to pick a feature set from the 9 features. This chapter doesn't aim to assess the classification performance of the chosen features. The features were only picked as sample EEG features for scalability tests so others may have been selected for this study.

3.5.3 Performing EEG feature selection with exhaustive search

We evaluate each of the 511 possible feature combinations to select the best feature combination for our classification problem. Figure 3.1 summarizes the feature evaluation steps. For simplicity, we choose KNN as classifier but the same principle applies to other classifiers. We then implement this algorithm in 4 steps in MapReduce:

1. Map: Extract the segments of interest from the original EEG files and compute all features for each of the segments
2. Reduce: Build one dataset per feature combination
3. Map: Train the classifier and assess its performance for each feature set
4. Reduce: Choose the feature set that maximizes mean accuracy (for all classes).

Details on the classifier and EEG segments of interest are found in Section 3.6.1.

[6] the mNNC value increases in seizures and decreases in sleep or encephalopathies

3.6 Experiments

This section describes the experiments performed and their setup. Table 3.1 summarises the hardware and software properties of the experimental servers.

3.6.1 Details on EEG classification

EEG labeling hinges on properties such as sequence type and patient age so feature selection can be done only on segments of similar properties. Only eyes-closed segments from adult EEGs are used in this chapter. The feature selection principle is unchanged for other age groups and segment types. We use KNN as a classifier. We assess a feature set's performance by the mean classification accuracy (mean of the accuracy for all classes) and run 3 rounds of the Shuffle and Split cross-validation, with 30% of the data used as training set per iteration, to reduce overfitting and minimize the prediction error. We have 3 EEG classes: normal, normal but for increased β wave (often due to medication) and abnormal.

3.6.2 Dataset description

We use a dataset of 2500 EEGs for the experiments. All EEGs in the dataset were recorded on patients in a hospital setting at the Medisch Spectrum Twente (Enschede, The Netherlands), following the International 10/20 System with Ag/AgCl electrodes and using a common average reference. This amount of data is about 30% of the EEG data collected monthly[7] in the Netherlands and about 2 years of data from the local hospital (Medisch Spectrum Twente(MST), Enschede, The Netherlands). Only the 19 channels common to all EEGs are kept for calculations, with each channel sampled at 250Hz. All 9 features from Section 3.5.2 and mNNC are computed on the whole dataset (hereafter named dataset 1-Table 3.1(a)) to check the scalability of feature computation. To test feature selection by exhaustive search, we use a subset of 1000 files from dataset 1 for which the class label is known precisely (hereafter named dataset 2). The EEGs in both datasets predominantly represent standard EEGs (15 to 40 minutes' EEGs) i.e the most common EEGs in clinical practice (91.6% of the EEGs recorded per year at the local hospital).

[7] and about a third of the annual Dutch data in filesize

Dataset	Number of files	Total size of files	Minimum EEG duration	Maximum EEG duration	Number of files of duration <15mn	15 to 40 mn	40mn to 1h	1 to 2h	>2h	Number of values
dataset 1 (feature computation only)	2500	46.51GB	10s	3h 9mn	204 (7.4% of files)	2201 (79.5% of files)	90 (3.25% of files)	253 (9.14% of files)	19 (0.69% of files)	578,648,474,500
dataset 2 for classification subset of dataset1)	1000	16.06GB	10s	2h 8mn 50s	73 (5.6% of files)	909 (69.9% of files)	33 (2.54% of files)	35 (2.69% of files)	1 (0.08% of files)	6,828,505,000

(a) Characteristics of experimental datasets

Server	OS	Software used	Processor	RAM	Number of nodes
Server for Parallel Python experiments	openSUSE 12.3 Milestone 2(x86-64) Kernel version 3.6.3-1-desktop	Python 2.7.3 with joblib 0.7d library scikit-learn 0.14	AMD Opteron® Processor 4226 (6 cores) 2 processors	64GB	1
Hadoop cluster	Ubuntu 12.04.2 LTS(x86-64) Kernel version 3.2.0-40-generic	Python 2.7.3 with scikit-learn 0.10 Hadoop streaming jar from Cloudera Hadoop CDH3u6	Intel® Xeon® CPU E3110@3.00GHz (2 cores) 1 processor	7.8GB	15

(b) Characteristics of the servers used in the experiments

Table 3.1: Server and EEG test file characteristics

3.6.3 Benchmarking the EEG exhaustive search feature selection

Setup We test EEG feature selection with python and with Hadoop Streaming. To speed up the python code, we use the joblib library to parallelize parts of the feature selection: features are computed EEG by EEG with several tasks running concurrently and several feature combinations are tested for classification at the same time. The number of jobs running concurrently is RAM-bound.

We selected Hadoop Streaming as Hadoop interface as we can write python code with it. This allows us to reuse most of the code from the python with joblib approach, thus easing the performance comparison between both approaches tested. There are 30 available map slots in the Hadoop cluster (2 maps per node) so that up to 30 maps run at the same time until the Hadoop map jobs are done. Similarly there are 30 possible reduce slots. Unless otherwise stated, we run 2 maps per node for the Hadoop Streaming jobs. We compute all features over windows of 1800 ms in both Hadoop and Python approaches.

1800 ms of EEG data equals 450 points per channel with the standard frequency of EEG signal i.e 250Hz and about 9 eye blink artifacts (shortest known EEG events).

Experiment 1: Feature computation

In the first set of experiments, we only perform the first step of feature selection (described in Section 3.5.3), i.e EEG segment extraction and feature computation, on part or all of dataset 1. For each experiment, execution times are recorded. Figures 3.3(a) and 3.2 were obtained using all of dataset 1. Figures 3.2 and 3.3(b) were obtained with the server configurations shown in Table 3.1(b). Figure 3.3(a) explores the evolution of feature computation times when the number of cores of the Python server is made to vary. Feature computation execution times grow linearly with the size of processed files for both Hadoop and Python solutions (Figure 3.3(b)) but the Python execution times grow 4.5 times faster than the Hadoop ones. Therefore, feature extraction with Hadoop is especially beneficial for large files and scales to a national scale amount of data. Based on the interpolations of Figure 3.3(b), extracting the 10 features from Section 3.5.2 for the whole annual Dutch EEG data(i.e 167GB-Table 1.1) would take about 11 hours and 7 minutes with Hadoop compared to more than 2 days with Python. The Python execution time decreases exponentially with the number of active cores/CPUs (Figure 3.3(a)) but an infinite number of CPUs would be needed to reach the same performance as Hadoop!

Experiment 2: Brute-force classification and feature selection

Experiments described in this section all use dataset 2 (see Section 3.6.2 and Table 3.1(a) for details) and test the time it takes to assess the classification performance of all possible 511 feature combinations [8]. 253295 EEG segments are extracted from dataset 2, i.e 113,982,750 values or 1.67% of the total values in the original files. Table 3.2 summarises the results of implementing the feature selection algorithm described in Section 3.5.3 with Hadoop Streaming and Python. Due to recurrent memory errors, only 154 feature combinations out of 511 (30.14%) were tested for classification with Python. The execution times for Python classification in Table 3.2 are estimates based on available data. Insufficient RAM per Hadoop node led to all 511 combinations being tested with 37 successive jobs [9] instead of one so that only 1 map would run per node and not 2. The current implementation is clearly subpar as map slots become available

[8] all features except nearest neighbor synchronization

[9] 36 testing 14 combinations at a time and 1 testing 7 combinations at a time

Table 3.2: Execution times for whole feature selection process on dataset 2 and each of its steps

		Segment extraction& feature computation only	Feature computation and formatting for classification	Classification only	Complete feature selection selection
Execution time	Hadoop streaming	30.35min	1h7min20s	32h25min52s[9]	33h33min12s
	parallel Python	97.9min	97.9min	estimated lower bound: 11 days 47min[10] estimated upper bound: 12 days 14h34min	estimated lower bound: 11days2h25min estimated upper bound: 12 days 16h2 min

as the job runs but are unusable until the job ends and the next starts. This is however easily fixed, with the right user privileges, by setting the maximum number of maps per node to 15 so that at any time only one map runs per node: all 511 classifications can then run in a single Hadoop job. Table 3.2 shows that even this suboptimal solution evaluates the classification performance of all feature sets faster than Python. The gap in classification execution times between Hadoop and Python widens with the size of datasets to classify (Figure 3.4(a)). For very small datasets (33 training and 67 test points), Python outperforms Hadoop slightly (1.82 minutes for Python and 2.4 minutes with Hadoop to test all 511 combinations). Hadoop has overall a clear edge over Python as dataset size rises: the classification runs about 64.76 times faster on Hadoop. Classifying dataset 2's sequences, even in suboptimal conditions with Hadoop, runs 29.9 to 34.16[10] times faster than with Python (see Table 3.2). So Hadoop is more suited for large datasets' classification. Hadoop also scales linearly with the size of classification input files[11] (Figure 3.4(b)) and handles feature dimensions' increase better than Python (about 2 orders of magnitude faster than Python (Figure 3.2)).

3.6.4 Discussion

The experiments (Section 3.6.3) show Hadoop as a scalable and promising solution to process EEGs if the task at hand is parallelizable (eg feature computation) even if it is CPU-intensive and RAM-bound (classification with all possible feature combinations). It goes to prove that a cluster of commodity hard-

[10] compared to the estimated upper and lower bounds for the Python job respectively

[11] files obtained by extracting all eyes closed segments from the original EDF+ files and applying each of the 9 tested features on the extracted segments

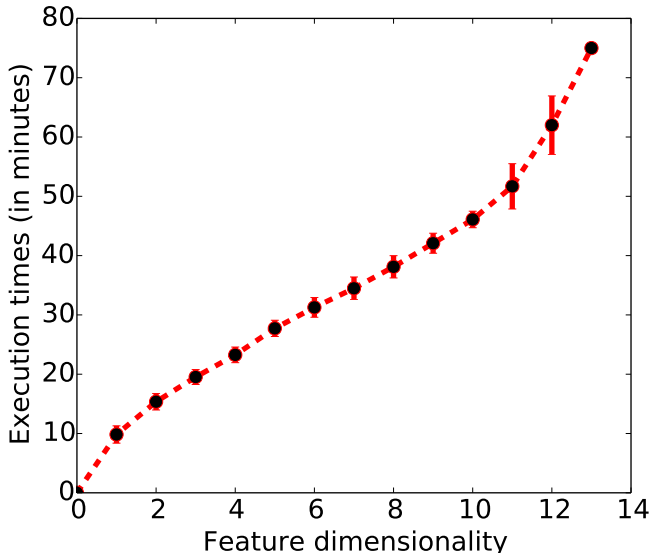
[15] Result of 37 successive jobs instead of only one job testing all 511 combinations

[16] estimates based on data available

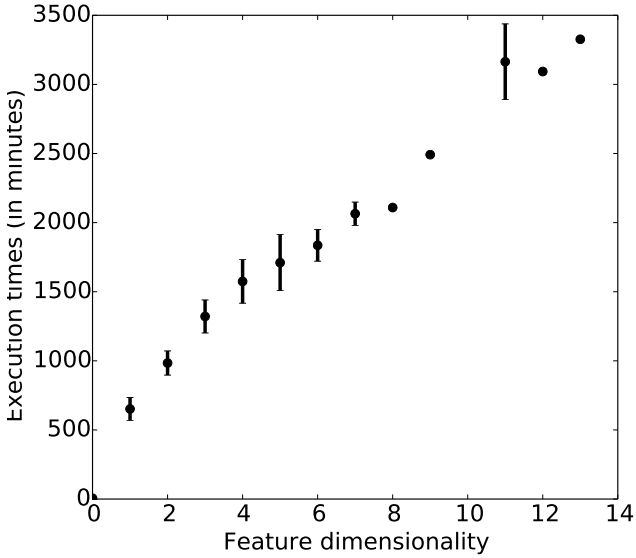
ware (15 machines with Dual core processors and only 7.8GB of RAM here) is better at processing complex data than a single highly specialized powerful server if the task is a series of (semi-)independent steps that can run in parallel. Hadoop has also been shown to be able to process a national scale amount of data with a quite small number of cluster machines. This is also a rather cheap solution: a cluster like the experimental one costs 10000 to 20000 euros i.e 1000-1500 euros per machine as compared to above 3000 euros per machine for the type of server used in the Python experiments. Owning a Hadoop cluster is in theory not needed as web services like Amazon Elastic Map Reduce (EMR) offer access to Hadoop clusters tailored for diverse processing needs. This is not doable, though, given the sensitivity of medical data. And we can boost the Hadoop performance further by optimizing the code we wrote by mostly reusing the Python one, via for example, changing the Hadoop configuration parameters to solve memory issues or using other Hadoop Python frameworks like mrjob or Dumbo that don't require map/reduce inputs and outputs to be strings passed via stdin/stdout and should thus need less processing RAM or using machine-learning algorithms optimized for the platform (Mahout library).

3.7 Conclusions

Hadoop is a promising solution for EEG storage and processing. Computation times for complex parallelizable machine-learning algorithms are notably reduced compared to more traditional means of computation and become manageable. The gain in computation times grows with data amount to process, Hadoop scaling easily with national scale data. So it would seem that it is better to process data with many commodity machines rather than with one extremely powerful server, when the processing task is parallelizable. In future, we would like to extend this work to other medical data types such as MRI or CT and study how to integrate data from computations run on diverse types of medical data (eg MRI and EEG). We would also like to run more tests on medical data querying (especially natural language querying). And Hadoop data security also needs to be explored further.

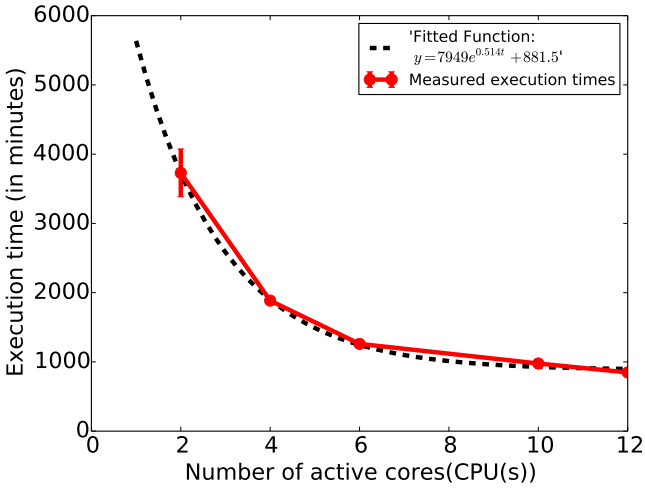


(a) Hadoop solution

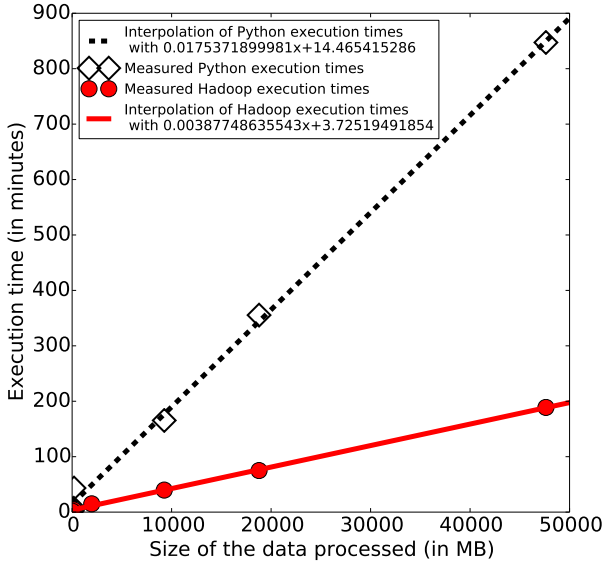


(b) parallel Python solution

Figure 3.2: Relationship between feature dimensionality and features' computation execution times

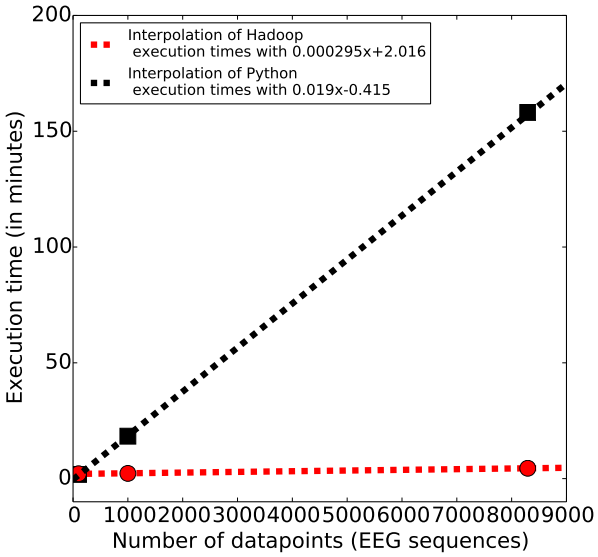


(a) Evolution of features' computation execution times (parallel Python) with number of active cores

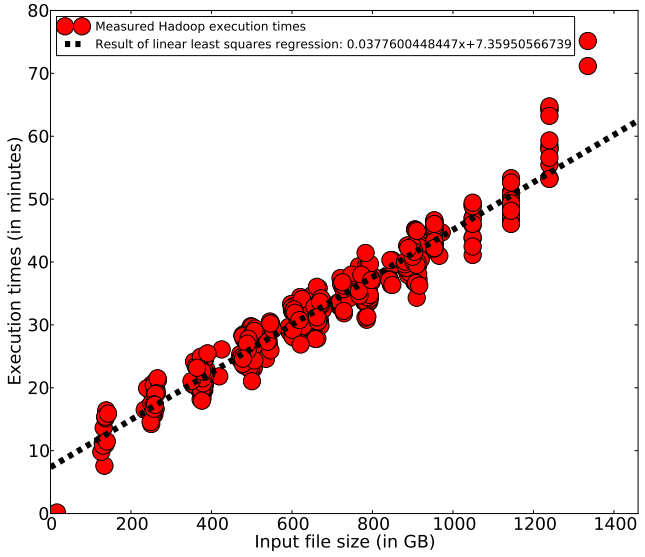


(b) Variation of feature computation execution times with EEG file size

Figure 3.3: Relationship between features' computation execution times and various parameters



(a) Variation of classification execution times with number of data points (test and training sets)



(b) Evolution of Hadoop classification execution times with classification input file size (1 file per combination)

Figure 3.4: Relationship between classification computation execution times and various parameters

CHAPTER 4

Evidence combination for incremental decision-making processes

The contents of this chapter have been submitted to the Knowledge and Information Systems (KAIS) journal.

4.1 Introduction

Reaching an accurate diagnosis as soon as possible is key to treating patients' ailments effectively. The case of teenager Rory Staunton who died of sepsis a few days after having been diagnosed with a benign flu at the ER and sent back home illustrates how critical it is to reach a timely accurate diagnosis [1].

Though quite extreme, Rory Staunton's case is not an isolated case of misdiagnosis and is just one particularly striking example of the many errors of diagnosis that occur in the healthcare system. In fact, the prevalence of misdiagnoses is estimated to be up to 15% in most areas of medicine [2] and a study of physician-reported diagnosis errors [3] finds that 28% of the misdiagnoses are major (i.e resulting in death, permanent disability, or near life-threatening event) and 41% moderate (i.e resulting in short-term morbidity, increased length of stay, higher level of care or invasive procedure). Even common conditions such as pneumonia, asthma or breast cancer are routinely being misdiagnosed especially when the presenting symptoms are atypical [5, 6]. Missed diagnoses alone account for 40,000 to 80,000 preventable deaths annually in the US [4].

Not only is reaching a correct diagnosis quite a challenge but the process leading to a reliable diagnosis is often rather lengthy as it may involve many patient consultations and referrals to other clinicians as well as various tests and scans.

The study of physician-reported errors of diagnosis cited earlier [3] also finds that 32% of the cases are due to clinician assessment errors. This figure coupled with the misdiagnosis prevalence figure suggests that the misdiagnosis problem is far from being an individual clinician's problem but rather a systemic problem. Or in the words of [55]:

> "Errors are rarely due to personal failings, inadequacies, and carelessness. Rather, they result from defects in the design and conditions of medical work that lead careful, competent, caring physicians and nurses to make mistakes that are often no different from the simple mistakes people make every day, but which can have devastating consequences for patients. Errors result from faulty systems not from faulty people, so it is the systems that must be fixed."

And since "to err is human", systems must be designed in such a way as to make errors hard to commit or to quote the Institute of Medicine landmark report on medical errors [56], "Human beings, in all lines of work, make errors. Errors can be prevented by designing systems that make it hard for people to do the wrong thing and easy for people to do the right thing".
As such, instead of focusing on assigning blame to physicians/nurses, which does little to fix systemic problems and only ensures that preventable errors are made again and again, it would be more beneficial to try and identify the factors that contribute to making it difficult to reach a correct diagnosis in a timely fashion or that lead to erroneous/delayed diagnoses. Some of these factors include the following:

1. *Only finite resources can be allocated to the diagnosis process*. Even with the best of intentions, a doctor can only devote a limited amount of time and energy to each patient under his/her care. Furthermore, to decrease costs and minimize patient discomfort, the number of tests performed to reach a diagnosis needs to be kept as low as possible. There is also only a fixed (small) number of specialists and doctors are encouraged to make as few referrals as possible. And obviously, even with the best will in the world, doctors, being human, have only a limited amount of memory and knowledge to draw on to make diagnoses.

2. *The diagnosis process is highly dependent on the accuracy of the initial diagnosis hypothesis*.
   The patient is at best an unreliable source of information: he/she may

give vague information or omit crucial clues that he/she feels are not significant. Moreover, patient history, which may shed a different light on some non-specific presenting symptoms, is usually fragmented and scattered across different healthcare institutions that don't necessarily share information between themselves. Therefore, the first patient consultation only provides incomplete and highly noisy information on which the clinician needs to rely to form his/her initial hypothesis and order the relevant tests and/or referrals required to unearth further relevant diagnostic clues and evidence.

3. *Finding the right clues and evidence for a diagnosis is comparable to searching for a needle in a haystack.* Patient consultations/referrals and medical tests generate a huge amount of data that may or may not contain the needed diagnostic clues (depending on whether the right hypotheses were tested for) and is mostly irrelevant for the diagnosis task at hand. There is at the same time too much and too little data available.

4. *Medical knowledge is fragmented.* Due to the sheer amount of medical knowledge accumulated through time, no single clinician can know all there is to know inevitably leading to a spread of expertise and knowledge between clinicians.

5. *The diagnosis process is fragmented*. The patient often has to consult several doctors and undergo several tests. This is a direct consequence of the fragmentation of knowledge and expertise driven by the massive amount of medical knowledge available.

As a result of these factors, the potential of communication breakdowns between healthcare agents and crucial information being lost in the process increases as does the likelihood of clinicians falling back on potentially harmful cognitive biases.

Rory Staunton's case [57, 1] is a case in point of how a *breakdown in communication between healthcare agents* can result in erroneous diagnosis and inadequate care. In Rory's case, because critical blood tests' results had not been communicated to the clinicians in charge and important observations by the pediatrician had gone missing from the charts, each of the parties involved in Rory's care had only access to fragments of information on his condition, each of which could be construed to result from something other than sepsis. Taken in conjunction, all of Rory's symptoms and tests pointed clearly to sepsis but the

flu diagnosis was not outlandish given the information available to the first ER practitioner at the time of diagnosis. This case perfectly exemplifies the situation described in the tale of the blind men and the elephant:[1] While the conclusions of the blind men might have been right individually, taken as a whole, they missed the target completely.

Rory's case also illustrates another source of diagnostic failure: *cognitive biases* [9, 58, 59]. Two biases were in play in Rory's case: *representativeness bias* and *premature closure*. The representativeness bias is the tendency for a clinician to look for prototypical manifestations of a condition, thus rejecting the possibility of a particular condition if the presenting symptoms are atypical or if the patient is not part of the stereotypical population in which the condition occurs. In Rory's case, the possibility of sepsis was not considered because sepsis rarely occurs in teenagers. Premature closure is the tendency for a clinician to decide on a diagnosis to the exclusion of others too soon in the process and before it has been fully verified by tests for example. In Rory's case, there was no indication that the attending clinicians had considered another diagnosis than flu. Cognitive biases are essentially reasoning shortcuts and heuristics that come into play when doctors try to cope with time and resources constraints. Cognitive biases are necessary and time-saving but may result in wrong, missed or delayed diagnoses.

In addition to the representativeness and premature closure biases, a few more biases may become problematic if applied indiscriminately: zebra retreat, availability and confirmation biases and diagnosis momentum. A clinician usually follows the well-known maxim "If you hear hoofbeats, think horses, not *zebras*[2]", i.e. a clinician tends to only consider the most common diagnoses that fit the symptoms exhibited by the patient. Failing to consider a zebra even when likely based on the clinical findings is called zebra retreat. Taken as a whole, zebras are not so uncommon: 8% of the US population (ie about 25 million) are estimated to be affected by one of the approximately 7000 zebras.

The *confirmation bias* can be especially harmful when associated with premature closure: it is the tendency for a clinician to look for the evidence, even not present, that supports his/her preferred diagnosis and dismiss the existing

[1] The story, which has several versions (see `http://en.wikipedia.org/wiki/Blind_men_and_an_elephant`), basically goes like this: some blind men or men in a dark room touch different parts of an elephant trying to figure out what they are touching. Depending on which part they touch (trunk, leg ,etc.), they come to completely different conclusions.

[2] Zebra is the medical slang for rare or surprising diagnosis. For examples of zebras, see the Medical Mysteries column in the Washington Post: `http://www.washingtonpost.com/linksets/medical-mysteries/2010/07/06/ABELr7D_linkset.html`.

evidence that disproves it. The confirmation bias can cause the clinician and patient to go on a wild-goose chase and delay the diagnosis especially if it intervenes while forming the initial diagnosis hypothesis since the whole process hinges on that initial hypothesis.

The *availability bias* and *diagnosis momentum* may be consequences of the fragmentation of expertise. The availability bias is the tendency of a clinician to reach for the most easily recalled diagnosis that fits the clinical findings, whether the clinician recalls that diagnosis because he has more expertise on it or because he has recently encountered it. Diagnosis momentum is the fact for a diagnosis to stick in particular because it keeps being passed on by all the agents and intermediaries involved in the diagnosis process. Diagnosis momentum also makes reaching a correct diagnosis during the initial patient consultation critical.

We contend that, to obviate or at least mitigate the aforementioned factors leading to misdiagnosis, different forms of computer-support could be used to assist clinicians in their decision-making task. One form of computer-support is *(semi-)automatic interpretation* of tests and scans, such as the semi-automated EEG interpretation performed by [60, 54]. A different form of computer-support, which is the focus of this chapter, is *evidence combination*. We view medical diagnosis primarily as an incremental process where at each point in time, there is an intermediary diagnosis based on 'what is known so far': symptoms and clinical evidence from consultations, tests/measurements/scans, interpretations thereof by experts, second opinions/feedback of experts on other experts' interpretations/conclusions, etc. Each interpretation, opinion, or feedback is a piece of evidence that is combined to produce a well-weighted intermediary diagnosis.

4.1.1 Contribution

The model presented in this chapter

1. provides a combined diagnosis constructed from all known evidence and opinions known so far at a point in time,
2. is based on Dempster-Shafer theory,
3. allows the inclusion of evidence that stems from the processing of historic evidence found in electronic patient records and usually not or insufficiently considered with the help of computers,

4. is open to including the outputs of computer-based interpretation tools as evidence,

5. allows a clinician to take into account more alternatives so as to notify him/her of rare diseases becoming sufficiently likely to warrant consideration,

6. can incorporate meta-evidence, i.e., feedback from one clinician on the diagnosis of another, and

7. protects him/her against ill-advised cognitive biases [9, 58, 59, 61]

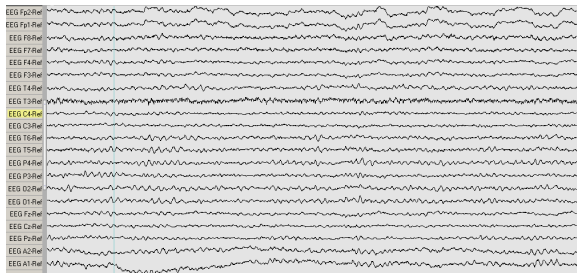
4.1.2 Examples

To illustrate the potential of computer-support through evidence combination with these properties, consider the following two examples.

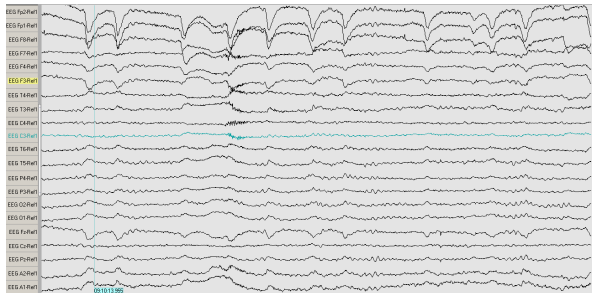
Example 1: Toothbrush case
A suspicious sequence is detected in the EEG recording of an ICU patient (see Figure 4.2). Several clinicians debate but they can't agree on a diagnosis based on this sequence: opinions are split between epilepsy and artifact. A few clinicians (2%) think it is something else, i.e., unknown. Subsequently, the video recorded simultaneously with the suspicious EEG sequence is reviewed. It shows without a doubt that the sequence is actually an artifact due to the patient brushing his teeth. Figure 4.3 shows a timeline of events for the toothbrush case.

Example 2: Hemochromatosis case
This example is a real case reported in a Washington Post article in the Medical Mysteries section (see [62]).

After an initial set of seemingly unrelated complaints and symptoms (blurry and rapidly decreasing vision, increased sleepiness, fatigue, high blood sugar level leading to a diabetes type I diagnosis), the patient lands in the ER with symptoms such as severe confusion and disorientation, internal bleeding and liver cirrhosis. Tests rule out the possibility of an infection or of hepatitis C and the ER doctors conclude that the symptoms exhibited by the patient result from a combination of diabetes type I and severe alcoholism (some symptoms being seen as signs of alcohol withdrawal). However, both the patient and his family deny the alcoholism especially since he hadn't drunk any alcohol in the two weeks before the ER visit as a result of fatigue. Moreover, some tests,



(a) Eyes closed



(b) Eyes open

Figure 4.1: Normal EEGs in different contexts

undisclosed by the ER personnel to the patient at that point, show an alcohol level of 0g/L and extremely high blood iron levels.

Unconvinced by the diagnosis given at the ER, the patient, with the help of a pathologist friend, researches possible explanations for the complaints that landed him in the ER. Hemochromatosis, a disease found while perusing a medical textbook,[3] appears a very likely possibility to him and after some tests (level of iron in the blood and genetic test), the hemochromatosis diagnosis is definitively confirmed. Figure 4.4 shows a timeline of events for the hemochro-

[3]genetic disease that causes the body to absorb and store excessive amounts iron, resulting in organ damage

Figure 4.2: EEG of a toothbrush artifact

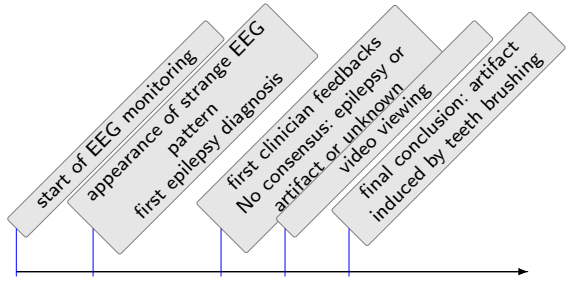


Figure 4.3: Chronology of events for the toothbrush case

matosis case.

These examples illustrate several things. First, medical tests and scans interpretation such as, for instance, EEG interpretation, are *inherently uncertain*. As [23] puts it: "there is an element of science and an element of art in a good EEG

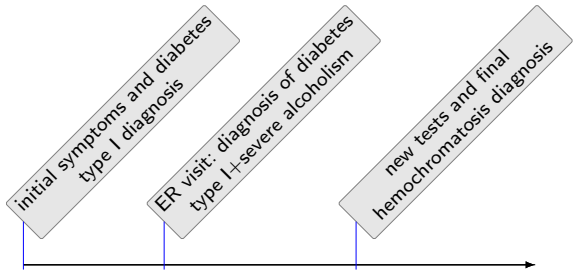


Figure 4.4: Chronology of events for the hemochromatosis case

interpretation" (p.167). The uncertainty of the interpretation can stem from the massive amount of data whose perusal is required to make a conclusion. For instance, the interpretation of a routine 20-minute EEG, usually done visually by a trained neurologist, requires the perusal of large amounts of data[4] that contain age-dependent, context-dependent and non-specific patterns. Uncertainty can also stem from the fact data patterns have no standard definition making the interpretation process not reproducible. For instance, a study by [30] showed that, even when done by one single clinician at two different points in time, markings on EEG recordings for patterns such as epileptiform discharges may not be identical.

Computer-based interpretation tools may help make the data interpretation process more reproducible. Crucially, interpretation tools, while producing uncertain results, provide a quantification of the result uncertainty as opposed to visual inspection by humans for which the uncertainty estimation is tricky. This means that it can effectively be included in the evidence combination. (Semi)-automated interpretation tools could also provide cognitive aids and exploitable markings to clinicians, thus reducing their workload and their reliance on memory and freeing up enough time for them to focus properly on hard cases only. This would not only make the diagnosis process faster but also less prone to errors due to mistaken cognitive biases [59]. And while not substitutes for clinician input, computer-based interpretation tools show good accuracy: a study trying to predict cancer outcomes based on applying

[4] at least 109 A4 sheets of paper (1 sec of EEG being represented by at least 25mm on paper taken in landscape format) following the guidelines of the American Clinical Neurophysiology Society [15].

machine-learning algorithms on electronic administrative records finds such algorithms at least as accurate as a panel of 5 expert clinicians [63].

Second, crucial clues or evidence may be 'hidden' in the vast amounts of available data, eg the blood iron levels in the hemochromatosis case. As computers are able to process a larger amount of data than humans possibly can and faster than they do, computer-based data pre-processing may assist the clinician in *uncovering clues and evidence hidden in mountains of data* even in cases when the clinician does not even suspect that there is something to find and point to zebras when they become likely due to multiple clues occuring together. For example, in the hemochromatosis case, software matching clinical findings with possible diagnoses may have highlighted the conjunction of high iron blood levels, no alcohol in the blood, diabetes and cirrhosis and pointed out that the hemochromatosis diagnosis became likely with those findings and should be explored. This would effectively result in forcing clinicians to consider several alternatives thus reducing the incidence and impact of premature closure. Note that this does require the data be available in digital format, which is not as straightforward as it may seem. According to [8], 30% of data in medical records, laboratory, and surgery reports, is not digitized. And 90% of the data generated by healthcare providers is discarded, for example, almost all video recordings from surgery.

Third, often *unexpected circumstances may mislead* a clinician during the decision making process. Even when known with hindsight, it is hard to correctly trace back the process and properly reconsider the then known evidence. For example, it may transpire, after several EEGs have been recorded, that a set of electrodes used for one or more EEGs were faulty, which would mean that the diagnoses in which these EEGs were involved may need to be reconsidered. By storing data provenance, i.e., derivation lineage that represent which diagnosis is derived from what evidence taken from what data, the clinicians can be supported with batch-wise reconsideration of their diagnoses. And clinicians may only need to be notified if the unexpected event changes their diagnosis significantly.

Fourth, each step in the diagnosis process and all evidence leading to it should be *accessible for review*. In our example, the initial epilepsy diagnosis could be reviewed by several clinicians and modified significantly because of the video evidence. As such, a review of the EEG and its accompanying video helped override a faulty initial conclusion resulting from incomplete evidence and get to the correct conclusion. Computer-based evidence combination can assist the clinicians in properly incorporating new evidence as well as meta-evidence

(feedback) in the overall diagnosis thus insuring that available data is used to make a decision.

The remainder of this chapter is organized as follows. The medical diagnosis examples of Section 4.1.2 serve as running examples for further supporting our claims for evidence combination as well as for categorizing evidence into types (Section 4.2). We then give some background on the Dempster-Shafer theory underlying our model in Section 4.3. In Section 4.4, we explain how to represent evidence and the uncertainty inherent to it using the Dempster-Shafer framework. We then proceed to present our evidence combination model (Section 4.5) and validate it analytically (Section 4.7). We show how our model can be used in practice by working out the running examples as well as an important theoretical example from literature in detail (Section 4.6). Finally, Section 4.8 discusses how the model can be implemented by storing evidence in a probabilistic database which naturally supports uncertainty in data and maintaining lineage.

4.2 Categorization of evidence types for evidence combination

The examples given in Section 4.1.2 highlight the fact that the diagnosis process is an *incremental process*. New evidence modifies the state of knowledge: every step in Figures 4.3 and 4.4 introduces one or more pieces of evidence. By evidence, we mean the interpretation of a lab result or any other medical test or scan but not the raw data itself. What we call *evidence* is the new diagnosis that the clinician forms based on raw medical data (eg lab results or scans). The number of pieces of evidence is generally limited since clinicians tend to and need to focus on a limited set of alternatives. Note that computer-based interpretations of medical tests (eg lab tests and scans) and/or presenting symptoms are also considered to be evidence since they are an interpretation of raw medical data.

We distinguish three main types of evidence:

1. Evidence on already considered alternatives.
2. Evidence that introduces a new alternative.
3. Meta-evidence: Evidence on the reliability of other pieces of evidence.

The first type of evidence assigns a likelihood either to one or more existing alternatives or to part of one or more existing alternatives (e.g., new evidence supporting epilepsy while previously only evidence supporting both epilepsy and artifact existed). Special cases of this type of evidence include corroboration and rejection, i.e., positive or negative feedback from one clinician on the diagnosis of another. In our model, these are represented by a likelihood of 1 and 0, respectively, assigned to one particular alternative.

The second type of evidence occurs when new evidence, that may or may not support one or more previous conclusions, also points to a diagnosis hypothesis not previously considered. An example of this type of evidence occurs in in the hemochromatosis case when hemochromatosis becomes a possible diagnosis aside from the combination of diabetes and alcoholism diagnosed at the ER.

An example of the third type of evidence is the genetic test and blood test in the hemochromatosis case: the tests invalidated the ER conclusions, i.e., reduced their reliability. Another example is the video evidence, in the toothbrush case, that confirmed the suspicion of artifact in the toothbrush case and lead to the rejection of the epilepsy diagnosis.

Evidence has several characteristics, on top of having a type (as explained earlier in the section). These characteristics can be summarized as follows:

- Evidence is uncertain and depends, for example, on the reliability of its source. We therefore attribute a confidence score c to each piece of evidence to quantify its reliability. If no knowledge on source reliability is available, one needs to assume, by default, that all sources of evidence are equally reliable.

- It is crucial that a concrete record of the dependencies between evidences (i.e., evidence provenance) be kept to ensure that pieces of evidence are properly combined or re-considered at any time. The reason for this is that, while evidence obviously appears in certain order during the diagnosis process, the evidence that arises at a certain point in time may refer to other specific pieces of evidence that arose earlier, for instance, in case of corroboration or meta-evidence.

4.3 A brief introduction to the Dempster-Shafer model

The Dempster-Shafer theory is a mathematical theory of evidence and can be viewed, in a finite discrete space, as a generalization of the traditional probability theory where probabilities are assigned to sets and not to mutually exclusive singletons. So, whereas, in traditional probability theory, evidence has to be associated with only one event, the Dempster-Shafer theory makes it possible to assign the evidence to a set of events. The Dempster-Shafer theory is therefore useful when evidence is available for a set of possible events and not for each possible event within the set and collapses to traditional probability theory in the case where evidence is available for all possible events within a set.

Let $\Theta = \theta_1, \ldots, \theta_N$ be a finite set of possible hypotheses. Θ is called the *frame of discernment* in the Dempster-Shafer theory. Note that, according to the Dempster-Shafer theory, each element $\theta_i \in \Theta$ where $i = 1, \ldots, N$ doesn't have to be a singleton. For example, in the case of a clinician defining possible diagnosis with non-mutually exclusive diseases (for example migraine M, sinusitis S and labyrinthitis L), the frame of discernement Θ could be defined as: $\Theta = \{\varnothing, M, S, L, MS, ML, SL, MSL\} = 2^{\{M,S,L\}}$. As can also be seen in the previous example, the frame of discernment is usually defined as an exclusive[5] and exhaustive non-empty[6] set of possible alternatives.

The Dempster-Shafer theory defines three important functions on this frame of discernment: the basic probability assignment also called mass function (otherwise known as m), the belief function (denoted *Bel*) and the plausibility function (denoted as pl).

The *mass function* (or *basic probability assignment*) m is a function that assigns a value in $[0, 1]$ to every subset $\mathcal{A}$ of Θ (such that $\cup_{\mathcal{A} \subseteq \Theta} \mathcal{A} = \Theta$) and satisfies:

$$m(\varnothing) = 0$$
$$\sum_{\mathcal{A} \subseteq \Theta} m(\mathcal{A}) = 1$$

(This means, in practice, that the sum of all basic probability assignments of subsets of the frame of discernment Θ of whom some may overlap may be different from 1 and actually higher than 1. It simply means that some evidence is counted more than once). The basic probability assignment defined for a

[5] in the previous example, it would mean that only one of the elements of Θ is the true diagnosis

[6] it contains at least the empty set $\varnothing$

subset $\mathcal{A}$, $m(\mathcal{A})$ is actually the degree of belief that the variable of interest falls within interval $\mathcal{A}$. However, $m(\mathcal{A})$ gives no indication as to the degree of belief that the variable of interest falls within any of the subintervals of interval $\mathcal{A}$. Additional evidence is needed for that.

The *belief* and *plausibility* functions can be interpreted respectively as the lower and upper bounds for probabilities, with the actual probability associated with the considered subset of 2^{Θ} in between the belief and plausibility values for that subset . The *belief* function *Bel* assigns a value in $[0, 1]$ to every non-empty subset $\mathcal{B}$ of Θ and is defined as:

$$Bel(\mathcal{B}) = \sum_{\mathcal{A} \subseteq \mathcal{B}} m(\mathcal{A})$$

The *plausibility* function *Pl* assigns a value in $[0, 1]$ to all sets $\mathcal{A}$ whose intersection with the set of interest $\mathcal{B}$ is not empty:

$$Pl(\mathcal{B}) = \sum_{\mathcal{A} \cap \mathcal{B} \neq \varnothing} m(\mathcal{A})$$

Both the *belief* and *plausibility* functions are non-additive, which means that the sum of all belief values associated with values in 2^{Θ} is not required to be equal to 1 and similarly for plausibility. Furthermore, the *mass function* m can be defined using the *belief* function with:

$$m(\mathcal{B}) = \sum_{\mathcal{A} \subseteq \mathcal{B}} (-1)^{|B-A|} Bel(\mathcal{A})$$

where $|B - A|$ is the cardinality of the difference between sets A and B. And we can derive *plausibility* from *belief* with: $Pl(\mathcal{B}) = 1 - Bel(\overline{\mathcal{B}})$ where $\overline{\mathcal{B}}$ is the complement of set $\mathcal{B}$. If $Bel(\mathcal{B}) = Pl(\mathcal{B}) = m(\mathcal{B})$, then we have defined a probability in the classical sense of the term.

An underlying assumption in the Bayesian theory is the existence of an ultimate refinement, that is "a frame of discernment so fine that it encompasses all possible distinctions and admits no further refinement" [64, p. 119]. In other words, the Bayesian theory supposes that all possible worlds are known and defined. While such an ultimate refinement would be conceptually convenient, it is also unrealistic as, in most real world applications, existing possible worlds for the system are discovered as we go and more evidence is gathered. In contrast, the Dempster-Shafer theory allows for ignorance, does away with

the ultimate refinement hypothesis and instead defines frame refinements and coarsenings. According Shafer [64, p.120], a frame of discernment is defined as "a set of possibilities that one does recognize on the basis of knowledge that one does have — or at least on the basis of distinctions that one actually draws and assumptions that one actually makes". In other words, the frame of discernment reflects the state of knowledge at a given point in time so it is quite normal, in practice, to begin by defining a coarse frame of discernment and then refine it (that is split sets defined in the initial frame of discernment into finer subsets) as more knowledge is accumulated. The existence of such a possibility of refinement is what will allow us to perform user feedback that actually adds new alternatives (see Section 4.5.2 for more details). For a more detailed and formal definition of frames of discernment, frame compatibility and frame refinements and coarsenings, we refer to chapter 6 in [64].

The Dempster-Shafer theory also provides means to combine evidence obtained from multiple sources, that provide different assessments for the frame of discernment and are supposed independent from each other. One such combination rule is the Dempster combination rule, which can be defined by (given two sources denoted 1 and 2):

$$m_{12}(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K} \quad \text{when } A \neq \varnothing$$
$$\text{where } K = \sum_{B \cap C = \varnothing} m_1(B) m_2(C)$$
$$m_{12}(\varnothing) = 0$$

The denominator in Dempster's rule is a normalization factor and has the effect of completely ignoring conflict and attributing all probability masses associated with conflict to the null set.

According [65], this omission of conflict may lead to some counterintuitive results as in the following example (and hereafter refered to as Zadeh's example). A patient is seen by two physicians for troubling neurological symptoms. The first physician gives a diagnosis of meningitis with an associated probability of 0.99 while admitting the possibility of a brain tumour with an associated probability of 0.01. The second physician believes the patient has a concussion with a probability of 0.99 or a brain tumour with a probability of 0.01. If we use the Dempster's combination rule with the available data, we get

$m(\{\text{brain tumour}\}) = \mathit{Bel}(\{\text{brain tumour}\}) = 1$. This result would imply that the most likely diagnosis is actually the one that both physicians find extremely unlikely.

Furthermore, the Dempster combination as well combination rules derived from it, such as Yager's combination rule and Zhang's center combination rule to name a few, suppose that all sources of evidence are equi-reliable. In our application, we suppose this is not the case and that a reliability score w_i — between 0 and 1 — is associated with each user giving feedback (how this reliability score is determined is beyond the scope of this chapter, see Section 4.5.3). One way of taking into account the difference of reliability between sources of evidence would be to use the mixing (or averaging) rule described in [66]:

$$m_{1\ldots n}(A) = \frac{1}{W}\sum_{i=1}^{n} w_i m_i(A) \qquad \text{where } W = \sum_{i=1}^{n} w_i$$

where n is the number of sources, w_i the weight associated with the i-th source and m_i the mass function associated with the i-th source. For more details on the Dempster-Shafer theory and the evidence combination rules, see [64], [66], [67] and [68].

4.4 Representation of uncertain evidence

Ideally, the uncertainty surrounding evidence is precisely known, but in practice it is often incomplete, coarsely known or completely missing. Therefore, evidence needs to be represented under various circumstances:

1. *Exact evidence likelihood values available*. For example, "the EEG of the patient points to epilepsy with confidence 0.8 and to an artifact with confidence 0.2."
2. *Missing likelihood values* For example, "the EEG of the patient shows an epileptic seizure or an artifact."
3. *Imprecise likelihood values*. For example, "the EEG of the patient shows epilepsy with a confidence of at least 0.7."
4. *Coarse likelihood values*. For example, "the EEG of the patient shows epilepsy or an artifact with confidence 0.8 or a normal pattern with confidence 0.2"

Of particular interest for our application, are the cases where evidence likelihood values are exactly known and where likelihood values are missing (by far the most common case in our application). We represent piece of evidence i and its associated uncertainty with (a) a mass function m_i assigning likelihood values to several alternatives, combined with (b) a weight w_i representing the reliability of the evidence (relative to other pieces of evidence). For example,

$$m_i(S) = \begin{cases} 0.8 & \text{if } S = \{\textsf{epilepsy}\} \\ 0.2 & \text{if } S = \{\textsf{artifact}\} \\ 0 & \text{otherwise} \end{cases}$$

represents evidence that the EEG points to epilepsy with confidence 0.8 and to an artifact with confidence 0.2. This is a case where exact likelihood values are available: all mass is comprised in singletons, i.e., sets containing a single label.

If no likelihood values are known, the mass can be assigned to a set containing multiple labels. For example,

$$m_i(S) = \begin{cases} 1 & \text{if } S = \{\textsf{epilepsy}, \textsf{artifact}\} \\ 0 & \text{otherwise} \end{cases}$$

represents evidence that the EEG points to epilepsy or an artifact.

Note that in the verbal expression of such evidence, one often does not mention the possibility that it could be entirely something else. We make this explicit in our model by introducing the explicit label '**other**'. This label represents all other diseases or conclusions not considered (yet). This allows a likelihood to be assigned to this label. For example,

$$m_i(S) = \begin{cases} 0.8 & \text{if } S = \{\textsf{epilepsy}, \textsf{artifact}\} \\ 0.2 & \text{if } S = \{\textsf{other}\} \end{cases}$$

represents the evidence that the EEG points to epilepsy or an artifact with confidence 0.8, but that one keeps open the possibility that it could be entirely something else with a confidence of 0.2. This conclusion can, for example, be drawn from circumstances where one estimates that the reliability of the sources is not perfect, but 80%. Note that with the inclusion of the explicit label **other**, there is no need for an 'otherwise'; the mass function m_i representing a piece of evidence is always complete.

In the sequel, we will consistently use the term *label* and symbol a for a single interpretation, such as epilepsy or artifact, and the term *alternative* and symbol A for a set of labels, such as $\{\text{epilepsy}, \text{artifact}\}$. We denote with $\mathcal{L}$ the set of all considered labels; $\mathcal{L} = \{\text{epilepsy}, \text{artifact}, \text{other}\}$ in the example above. Therefore, the frame of discernment is $\mathcal{F} = 2^{\mathcal{L}}$. In the example above,

$$\mathcal{F} = \{\emptyset, \{\text{epilepsy}\}, \{\text{artifact}\}, \{\text{other}\}, \{\text{epilepsy}, \text{artifact}\}, \{\text{epilepsy}, \text{other}\}, \{\text{artifact}, \text{other}\}, \{\text{epilepsy}, \text{artifact}, \text{other}\}\}$$

4.5 Evidence combination model

4.5.1 Core of the model: the mixing rule

In Section 4.3, we introduced a basic probability assignments' combination rule, called the mixing (or averaging) rule defined as:

$$(\forall A \in \mathcal{F}) \quad m_{1\ldots n}(A) = \frac{1}{W} \sum_{i=1}^{n} w_i m_i(A) \qquad \text{with} \qquad W = \sum_{i=1}^{n} w_i$$

where n is the number of evidences, w_i the weight associated with the i-th piece of evidence, W the normalization factor being the sum of all weights, and m_i the mass function associated with the i-th piece of evidence.

We assume that a database actually contains both the individual pieces of evidence with all associated information as well as an aggregation of the evidence obtained from l previous sources. So, if we want to combine a new piece of evidence m_{l+1} with the l previous evidences, the total number of sources of evidence combined is $n = l + 1$. Also, the introduction of a new weight w_{l+1} updates the normalization factor $W' = W + w_{l+1}$. We apply the mixing rule as

follows:

$$
\begin{aligned}
(\forall A \in \mathcal{F}) \quad m_{1\ldots n}(A) &= \frac{1}{W'} \sum_{i=1}^{n} w_i m_i(A) \\
&= \frac{1}{W'} (\sum_{i=1}^{l} w_i m_i(A) + w_{l+1} m_{l+1}(A)) \quad since \quad n = l+1 \\
&= \frac{1}{W'} (W \frac{1}{W} \sum_{i=1}^{l} w_i m_i(A) + w_{l+1} m_{l+1}(A)) \\
&= \frac{W}{W'} m_{db}(A) + \frac{w_u}{W'} m_u(A)
\end{aligned}
$$

where $m_{db}(A)$ is the basic probability assignment associated with alternative A in the database, $m_u(A)$ is the basic probability assignment associated with alternative A by the user providing the new evidence and w_u the weight representing the reliability of this evidence.

Observe here that m_n representing the new combined diagnosis of all $n = l+1$ evidences, can be calculated incrementally in terms of m_{db}, m_u, and w_u. Also, defined in this way, the combination rule is trivially associative and commutative as well as idempotent.

4.5.2 Basic operations of the model

We model all types of evidence with three atomic operations. We present the third atomic operation in two separate cases: in practice there are two different types of evidence that can be handled with one atomic operation, i.e., 3a and 3b are formally the same operation.

1 Adding a (weighted) basic probability assignment m_u with weight w_u due to a new piece of evidence.

2 Updating the weights associated with one or more previously given evidences $J \subseteq [1..n]$.

3a Refining the frame of discernment by splitting a known label a into multiple more refined ones.

3b Refining the frame of discernment by adding a new label for something previously not considered.

Notation	Meaning
a	existing label, e.g., epilepsy.
A	existing alternative, e.g., {epilepsy, artifact}.
n	number of evidence sources prior to new evidence being added.
m_u	new evidence represented as a basic probability assignment. dom$(m_u) = \mathcal{F}$
m_{old}	stored combined basic probability assignment derived from $m_1, \ldots, m_n$ prior to new evidence
m_{new}	combined basic probability assignment after taking new evidence into account
w_u	weight associated with m_u
W	normalization factor prior to new evidence being added, $W = \sum_{i=1}^{n} w_i$.
W'	normalization factor after taking new evidence into account $W' = W + w_u$.
$J \subseteq [1..n]$	set of indices corresponding to evidence sources for which the weights must be updated because of new evidence
w'_j	new weight due to new evidence $(j \in J)$
$\mathcal{L}$	set of all considered labels
$\mathcal{F}$	frame of discernment, i.e, set of all considered alternatives, $\mathcal{F} = 2^{\mathcal{L}}$

Table 4.1: List of notations

The notations used in this section can be found in Table 4.1. Section 4.5.2 describes how to determine which atomic operation to use.

As explained earlier, we assume that the database also contains the mass function m_{old} representing the combined diagnosis of all n previous evidences. Through the atomic operations formulas derived below, we aim to recalculate a new combined mass function incrementally, i.e., to define m_{new} in terms of m_{old} and the basic probability assignment $m_u(A)$ associated with the new evidence.

Adding a (weighted) basic probability assignment

This operation is used when evidence is added without any change in the weights associated with previous evidence sources. The number of sources with the addition of new evidence becomes $n + 1$. The new normalization fac-

tor is $W' = W + w_u$. Applying the mixing rule gives us:

$$(\forall A \in \mathcal{F}) \quad m_{new}(A) = \frac{1}{W'}(\sum_{i=1}^{n} w_i m_i(A) + w_u m_u(A))$$
$$= \frac{W}{W'} m_{old}(A) + \frac{w_u}{W'} m_u(A)$$

Updating weights

This operation is used when new evidence leads to updating one or more weights w_j associated with previously given evidence with a new weight w'_j (e.g., decreasing the weight associated with a previous evidence because the source of evidence has been discovered to be less reliable than previously thought, or altogether canceling a piece of evidence with setting $w'_j = 0$).

The new normalization factor can be defined as

$$W' = \sum_{j \notin J} w_j + \sum_{j \in J} w'_j$$
$$= W - \sum_{j \in J} w_j + \sum_{j \in J} w'_j$$
$$= W + \sum_{j \in J} (w'_j - w_j)$$

We denote the latter *normalization correction* term with $W_\Delta = \sum_{j \in W} (w'_j - w_j)$.

According to the mixing rule, we have:

$$(\forall A \in \mathcal{F}) \quad m_{old}(A) = \frac{1}{W} \sum_{i=1}^{n} w_i m_i(A)$$
$$= \frac{1}{W} \sum_{j \notin J} w_j m_j(A) + \frac{1}{W} \sum_{j \in J} w_j m_j(A)$$

The updated basic probability assignment for alternative A is obtained by us-

ing the mixing rule as follows:

$$
\begin{aligned}
&(\forall A \in \mathcal{F}) \\
&\quad m_{new}(A) = \frac{1}{W'}(\sum_{j \notin J} w_j m_j(A) + \sum_{j \in J} w'_j m_j(A)) \\
W' m_{new}(A) &= \sum_{j \notin J} w_j m_j(A) + \sum_{j \in J} w'_j m_j(A) \\
&= \sum_{j \notin J} w_j m_j(A) + \sum_{j \in J} w'_j m_j(A) + \sum_{j \in J} w_j m_j(A) - \sum_{j \in J} w_j m_j(A) \\
&= \sum_{j \notin J} w_j m_j(A) + \sum_{j \in J} w_j m_j(A) + \sum_{j \in J} w'_j m_j(A) - \sum_{j \in J} w_j m_j(A) \\
&= W m_{old}(A) + \sum_{j \in J} (w'_j - w_j) m_j(A) \\
m_{new}(A) &= \frac{W}{W'} m_{old}(A) + \frac{1}{W'} \sum_{j \in J} (w'_j - w_j) m_j(A)
\end{aligned}
$$

Updating the weights basically consists of canceling the terms in which the weights to be updated appear and then adding the newly weighted basic probability assignments. A full incremental calculation is not possible in this case, but one needs to revisit all evidences in the database for which the weight is updated. Usually, this remains rather limited.

Refining the frame of discernment: splitting a label

The label epilepsy actually represents a set of *epileptic syndromes* that differ by the specific features that are present. For example, benign rolandic epilepsy, childhood absence epilepsy, and juvenile myoclonic epilepsy are all particular cases of epilepsy. Suppose that in a diagnostic process, there have only been pieces of evidence where a confidence is assigned to alternatives that include the label epilepsy, but that a new piece of evidence points to, say, childhood absence epilepsy, or juvenile myoclonic epilepsy, how can we properly represent this new evidence and combine it with the existing ones?

Talking about measurement results, [67, p.38] says:

> When measurement results are considered, the basic probability number $m(A)$ can be interpreted as the degree of belief that

> the measurement result falls within interval A; but $m(A)$ does not provide any further evidence in support to the belief that the measurement result belongs to any of the various subintervals of A. This means that, if there is some additional evidence supporting the claim that the measurement result falls within a subinterval of A, say $B \subset A$, it must be expressed by another value $m(B)$.

The labels used in our model are very similar to concepts in description logic [69]. A classic example of a description logic definition is $\mathsf{Person} \equiv \mathsf{Male} \sqcup \mathsf{Female}$. It defines the concept of a person to be equivalent to either a male or a female. Important to note, is that this definition also states that the union of all possible males in the past, present, and future, and all possible females in the past, present, and future, will *exactly* give one the set of all possible persons in the past, present, and future. In other words, this definition truly *refines* the concept Person into two sub-concepts (it doesn't state that Male and Female are disjoint though).

We also apply this technique of refining concepts to our labels. Here we call it *splitting a label*. We may define

$$\begin{aligned}\mathsf{epilepsy} \equiv &\mathsf{benign\ rolandic\ epilepsy} \sqcup \mathsf{childhood\ absence\ epilepsy} \\ &\sqcup \mathsf{juvenile\ myoclonic\ epilepsy} \sqcup \mathsf{other\ epileptic\ syndromes}\end{aligned}$$

Note that the inclusion of a label other epileptic syndromes is necessary, because otherwise the equivalence doesn't hold. For brevity, we use the shorthands bre, cae, jme and oes in the sequel.

We can use this equivalence of concepts for refining our frame of discernment in a non-interfering manner: by replacing epilepsy with the four sub labels. Formally,

$$\begin{aligned}\mathcal{L}' =&(\mathcal{L} \backslash \{\mathsf{epilepsy}\}) \cup \{\mathsf{bre}, \mathsf{cae}, \mathsf{jme}, \mathsf{oes}\} \\ \mathcal{F}' =&2^{\mathcal{L}'}\end{aligned}$$

Furthermore, we need to adapt all existing pieces of evidence to the new frame of discernment. This is done by similarly replacing epilepsy in all pieces of evidence, i.e., whenever a mass function contains an alternative $A = \{\mathsf{epilepsy}, \ldots\}$, we replace A by $\{\mathsf{bre}, \mathsf{cae}, \mathsf{jme}, \mathsf{oes}, \ldots\}$. Assigned confidences and weights remain unchanged.

Observe that the old frame of discernment $\mathcal{F}$ is compatible with the new one $\mathcal{F}'$, because it is a proper refinement. Analogously, the thus constructed mass functions are proper refinements as well.

Typically, this atomic operation is triggered by the occurrence of evidence for a sub-label of an existing label, for example the occurrence of evidence on {bre which is a subset of $A = \{\text{epilepsy}, \ldots\} = \{\text{bre}, \text{cae}, \text{jme}, \text{oes}, \ldots\}$. In this case, two successive operations are performed. First, the operation of splitting the label is carried out. In a sense, this first step makes the frame of discernment and all existing evidence compatible with the refined nature of the new evidence. This first operation is then followed by performing the atomic operation of Section 4.5.2 to add the new (sub-label) evidence. The result is a mass function in which all the terms defined prior to the new evidence are updated (with atomic operation of Section 4.5.2) and a new term is added for the sub-label on which new evidence is given. In the previous example, this means that a new term $m(\text{bre})$ is defined while all the other terms existing prior to the new evidence on bre are updated.

Generally speaking, the atomic operation of splitting a label a into sub-labels $a_1, \ldots, a_m$ is defined with the following steps:

1. Define the equivalence $a \equiv a_1 \sqcup \ldots \sqcup a_m$.
2. Let $(\forall A \in \mathcal{F})$ *refine*$(A) = \begin{cases} (A \backslash \{a\}) \cup \{a_1, \ldots, a_m\} & \text{if } a \in A \\ A & \text{otherwise} \end{cases}$
3. Refine the frame of discernment $\mathcal{L}' =$ *refine*$(\mathcal{L})$; $\mathcal{F}' = 2^{\mathcal{L}'}$.
4. For every $i \in [1..n]$, we define a refined mass function $m' =$ *refine*(m) as

$$(\forall A' \in \mathcal{F}') \quad \begin{cases} m'(A') = m(A) & \text{if } \exists A \in \text{dom}(m) : A' = \textit{refine}(A) \\ 0 & \text{otherwise} \end{cases}$$

(Note that we have overloaded *refine* to work both on alternatives as well as mass functions).

Refining the frame of discernment: adding a new label

When a clinician makes a diagnosis, (s)he not only makes a diagnosis, but effectively excludes all other possible diagnoses. In this diagnosis, (s)he implicitly assigns a zero confidence to alternative {other}. The existence of the other label makes the frame of discernment as well as all mass functions exhaustive.

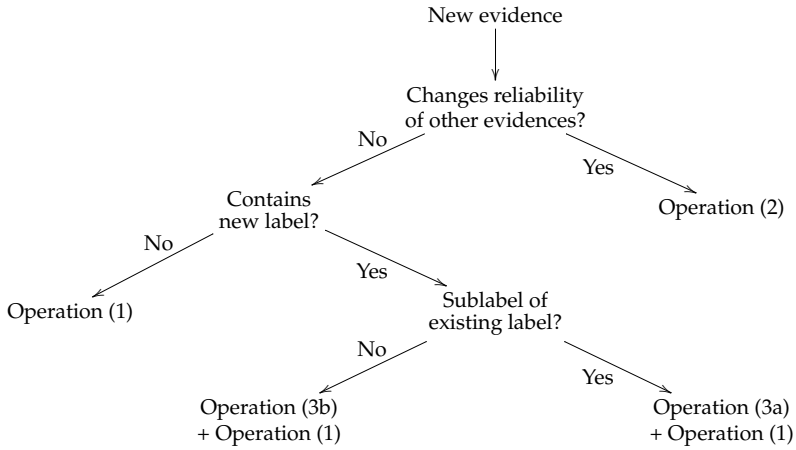


Figure 4.5: Decision tree for combining atomic operations to handle all types of evidence

It is the existence of the other label, however, that makes it possible to apparently "expand" the frame of discernment and add a new label a. For example, adding the hemochromatosis alternative after a diagnosis of diabetes combined with alcoholism has been reached in the hemochromatosis example. Because any new unknown label is already included in the other label, we can split it into a and a new other$'$ by defining other $\equiv a \sqcup$ other$'$ and applying the atomic operation of Section 4.5.2 (usually followed by the atomic operation of Section 4.5.2 since there is new evidence on a). Therefore, adding a new label is only a special case of refining the frame of discernment by splitting a label.

Deciding on which atomic operations to use

In Section 4.5.2, we introduced the atomic operations that are used to model the addition of new evidence. We here provide a decision tree that illustrates how the atomic operations need to be combined to handle all types of evidence.

4.5.3 Deciding on a weighting method

The setting of weights is not the purpose of this chapter. However, some ways to set the weights include defining rules to set weights (e.g., "clinician A is

twice as reliable as clinician B" or "video-based evidence supersedes EEG-based evidence") or deducing the weights by, for example, evaluating experts or sources of evidence through a set of calibrating questions for which the answer is known.

The mixing rule is a generalization of the averaging of probability distributions ([66]) also known as linear opinion pool. The linear opinion pool is widely used as a way to combine expert opinions in a probabilistic framework and several ways to set the weights in the linear opinion pool have been studied ([70, 71, 72, 73]). Similar strategies can be applied to set the weights for the mixing rule as well. One such strategy is the performance-based Cooke "classical" method . Cooke argues that using equal weights for all experts leads to a suboptimal solution as it doesn't evaluate the quality of each expert's opinion. Cooke suggests assigning the weight base on the performance of experts on an elicitation exercise based on "seeding variables", the "seeding variables" being quantities from the same area as the uncertain quantity of interest for which the true value is known to the one administering the exercise and not the experts. The experts may be asked to choose the probability bin in which they think the "seeding variable" they are given falls. Two scores are deduced from the experts performance: a calibration score which is the likelihood that the expert's answer corresponds to the known answer and an information/informativeness score that measures how concentrated the distribution given by expert is. Those two scores are then combined into a weight assigned to the expert. An expert that is "highly reliable" scores high on both calibration and informativeness.

Another way of determining appropriate weights is through data mining. At the end of a diagnostic and treatment process, the correct diagnosis is known. All evidence given in the process can then be evaluated based on its degree of correctness. By accumulating these evaluations for pieces of evidence given by a certain expert or a certain evidence source, one can determine an appropriate reliability score. Over time, one could determine a set of weights that is based on how accurate experts and sources actually are on average.

4.5.4 Rationale for using Dempster-Shafer framework instead of Bayesian framework

The Bayesian theory is a special case of the Dempster-Shafer evidence theory according to [64, Chp.1], with the Bayesian belief functions a subset of the Demspter-Shafer belief functions. The Dempster-Shafer theory is shown in [74]

to be more suited in cases of missing priors and ignorance. [64] tries to show through an example (Example 1.6, chapter 1, pages 23–24) that applying the Bayesian theory to cases of complete ignorance could lead to counter-intuitive results. In the example given by Shafer, the question is to know whether or not there is life around the star Sirius. And though some scientists may have evidence on this question, Shafer takes the point of view of the majority of people who profess complete ignorance on the subject and that

$$Bel(A) = \begin{cases} 0 & \text{if } A \neq \Theta \\ 1 & \text{if } A = \Theta \end{cases}$$

where $\Theta = \{\theta_1, \theta_2\}$, with θ_1 denoting the possibility that there is life on Sirius and θ_2 denoting the possibility that there is no life on Sirius.
He then considers a more refined set of possibilities $\Omega = \zeta_1, \zeta_2, \zeta_3$ where ζ_1 is the possibility that there is life on Sirius, ζ_2 the possibility that there are planets around Sirius but no life, and ζ_3 the possibility that there are not even planets around Sirius. The original frame of discernment Θ and the refined set Ω are related in that

$$\theta_1 = \zeta_1 \quad \text{and} \quad \theta_2 = \{\zeta_2, \zeta_3\}$$

which means that

$$Bel(A) = \begin{cases} 0 & \text{if } A \neq \Omega \\ 1 & \text{if } A = \Omega \end{cases}$$

So translating complete ignorance in the Dempster-Shafer framework is straightforward. Shafer goes on to try and show that it is difficult to specify consistent degrees of belief over Θ and Ω in the Bayesian framework when representing complete ignorance. Complete ignorance on Θ may be represented by $Bel(\{\theta_1\}) = Bel(\{\theta_2\}) = \frac{1}{2}$. On Ω, however, according to him, complete ignorance would mean that $Bel(\{\zeta_1\}) + Bel(\{\zeta_2\}) + Bel(\{\zeta_3\}) = 1$, hence $Bel(\{\zeta_1\}) = Bel(\{\zeta_2\}) = Bel(\{\zeta_3\}) = \frac{1}{3}$.
This yields

$$Bel(\{\zeta_1\}) = \frac{1}{3} \quad \text{and} \quad Bel(\{\zeta_2, \zeta_3\}) = \frac{2}{3}$$

These results are inconsistent with the ones found on Θ since $\{\theta_1\}$ and $\{\zeta_1\}$ have the same meaning as well as $\{\theta_2\}$ and $\{\zeta_2, \zeta_3\}$. However, this line of

reasoning is flawed. In fact, instead of considering the three possible events $\{\zeta_1, \zeta_2, \zeta_3\}$, one should consider four events. Let event A be "There is life on Sirius" and B the event "There are planets around Sirius". A and B are independent. Based on A and B, there are four events on Ω instead of three: $a = A \wedge B$, $b = A \wedge \neg B$, $c = \neg A \wedge \neg B$, $d = \neg A \wedge B$.

If $P(B) = \alpha$ then we know

$$P(a) = \frac{1}{2}\alpha \quad \text{and} \quad P(b) = \frac{1}{2}(1-\alpha) \quad \text{and} \quad P(c) = \frac{1}{2}(1-\alpha) \quad \text{and} \quad P(d) = \frac{1}{2}\alpha$$

Since $\{a, b\} = \theta_1$ and $\{c, d\} = \theta_2$, the solution obtained through the Bayesian method is still consistent and equivalent to the Dempster-Shafer solution. So when working with equivalent formulations, the solutions reached in both the Dempster-Shafer framework and the Bayesian framework are similar. However, the Bayesian framework calls for making assumptions (independence of A and B) and finding out some variables' values ($P(B)$) to reach a solution, when no assumptions or additional variable values besides what is already known are needed to reach a solution in the Dempster-Shafer framework. Reaching a solution in the Bayesian framework when no independence assumption can be made is more difficult.

[74] compare the use of the Bayesian theory and Dempster-Shafer theory to combine evidence from sensors. They conclude

> "Both methods for dealing with uncertainty yield similar results if based on equivalent formulations. [...][W]e believe that Bayesian theory is best suited to applications where there is no need to represent ignorance, where conditioning is easy to extract through probabilistic representation, and prior odds are available. Dempster-Shafer theory is a more natural representation for situations where uncertainty cannot be assigned probabilistically to a proposition or its complement and when conditioning effects are either impractical to measure separately from the event itself or a simple propositional refinement, and prior odds are not relevant."

In practice, in our case study (the diagnosis process) in particular, ignorance is frequent. And rare are the cases where strong assumptions such as variable independence can be made. Our case study — the diagnosic process— is therefore best represented in the Dempster-Shafer framework rather than the Bayesian framework, since ignorance is a mainstay of the diagnosis process.

Though there have been many studies that show how to successfully model meta-evidence by using Bayesian or Markov networks [75, 76], we think such models may be unsuitable for our application, because

- the case where new evidence leads to the addition of a new alternative cannot be represented with such networks, because such evidence is not easily represented in a graph
- it would be counter-productive to use two different models (Bayesian/Markov for positive or negative feedback and another model for other types of evidence such as the addition of a new alternative), when we can use one model (based on Dempster-Shafer theory) for all types of evidence.

4.5.5 Mixing rule versus Dempster combination rule

We use the mixing rule above in our model rather than combination rules such as Dempster's combination rule, Yager's combination rule or Zhang's combination rule, because it allows the combination of evidence coming from sources that may not be equally reliable.
Furthermore, as explained in [77], the classic Dempster combination rule assumes the following:

- the list of alternatives contained in the frame of discernment is an exclusive and exhaustive list of hypotheses,
- all the sources of evidence combined are independent and provide independent evidence, and
- all sources of evidence are homogeneous i.e. equally reliable

All three conditions required for the proper application of the Dempster combination rule do not hold for the medical diagnosis process. The sources' independence cannot be guaranteed as clinicians(sources) may consult each other while trying to come up with a diagnosis. The sources are not necessarily equally reliable, for instance, in our running toothbrush example, the video-based feedback is more reliable than the EEG interpretation. And finally, the frame of discernment may not necessarily be exhaustive as new alternatives may crop up during the diagnostic process (e.g., the hemochromatosis alternative was considered after the ER-visit by the patient and his pathology-friend in the hemochromatosis case).

4.6 Using the feedback model: some examples

In this section, we illustrate the usage of our model in practice by applying it to the two examples introduced in Section 4.1.2 and to Zadeh's canonical example (introduced in Section 4.3).

4.6.1 First example: the toothbrush case

Here we apply our model to the toothbrush example from Section 4.1.2. The chronology of events in this case can be found in Figure 4.3. After the start of EEG monitoring, the appearance of a strange EEG pattern was observed. This alone does not carry any evidence that points to a possible diagnosis.

Several clinicians, then, debate the issue but they do not reach a consensus. Most of the clinicians (98%) are split between epilepsy and artifact. The rest of the clinicians (2%) think it is something else altogether, in other words not epilepsy or artifact. The new evidence resulting from the debate can be represented with the mass function below.

$$\begin{aligned} \mathcal{L} =& \{\mathsf{epilepsy}, \mathsf{artifact}, \mathsf{other}_1\} \\ w_1 =& 1 \\ m_1(A) =& \begin{cases} 0.98 & \text{if } A = \{\mathsf{epilepsy}, \mathsf{artifact}\} \\ 0.02 & \text{if } A = \{\mathsf{other}_1\} \\ 0 & \text{if } A = \varnothing \end{cases} \end{aligned}$$

The next piece of evidence stems from watching and interpreting the video. The conclusion derived from the video is that the strange EEG sequence is an artifact. We may treat this evidence in two ways. It could be seen as new evidence which is much much more reliable than the evidence resulting from the debate, for example, with a weight of $w_2 = 100$. Or, we may interpret the evidence as including the meta-evidence that earlier diagnoses are totally erroneous, because they relied on a false assumption regarding the recording context, while in fact the patient was brushing his teeth. Let us work out the case where the new evidence is considered way more reliable than all previous

evidence.

$$\begin{aligned}\mathcal{L} =& \{\mathsf{epilepsy}, \mathsf{artifact}, \mathsf{other}_1\} \\ w_2 =& 100 \\ m_2(A) =& \begin{cases} 1 & \text{if } A = \{\mathsf{artifact}\} \\ 0 & \text{if } A = \{\mathsf{epilepsy}\} \\ 0 & \text{if } A = \{\mathsf{other}_1\} \\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

According to the mixing rule, we can now determine a combined diagnosis.

$$\begin{aligned}\mathcal{L} =& \{\mathsf{epilepsy}, \mathsf{artifact}, \mathsf{other}_1\} \\ W_{12} =& 101 \\ m_{12}(A) =& \begin{cases} 0 & \text{if } A = \{\mathsf{epilepsy}\} \\ 0.990099 & \text{if } A = \{\mathsf{artifact}\} \\ 0.009703 & \text{if } A = \{\mathsf{epilepsy}, \mathsf{artifact}\} \\ 0.000198 & \text{if } A = \{\mathsf{other}_1\} \\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

Because of the occurrence of singular alternatives $\{\mathsf{epilepsy}\}$ and $\{\mathsf{artifact}\}$ as well as a combined alternative $\{\mathsf{epilepsy}, \mathsf{artifact}\}$, the situation is not immediately clear. Here, the notions of belief and plausibility help to obtain a lower

and upper bound:

$$
\begin{aligned}
\mathit{Bel}(\{\mathsf{artifact}\}) &= \sum_{A \subseteq \{\mathsf{artifact}\}} m_{12}(A) \\
&= m_{12}(\varnothing) + m_{12}(\{\mathsf{artifact}\}) \;=\; 0.990099 \\
\mathit{Pl}(\{\mathsf{artifact}\}) &= \sum_{A \cap \{\mathsf{artifact}\} \neq \varnothing} m_{12}(A) \\
&= m_{12}(\{\mathsf{artifact}\}) + m_{12}(\{\mathsf{epilepsy}, \mathsf{artifact}\}) \;=\; 0.999802 \\
\mathit{Bel}(\{\mathsf{epilepsy}\}) &= \sum_{A \subseteq \{\mathsf{epilepsy}\}} m_{12}(A) \\
&= m_{12}(\varnothing) + m_{12}(\{\mathsf{epilepsy}\}) \;=\; 0 \\
\mathit{Pl}(\{\mathsf{epilepsy}\}) &= \sum_{A \cap \{\mathsf{epilepsy}\} \neq \varnothing} m_{12}(A) \\
&= m_{12}(\{\mathsf{epilepsy}\}) + m_{12}(\{\mathsf{epilepsy}, \mathsf{artifact}\}) \;=\; 0.009703
\end{aligned}
$$

In other words, the likelihood of an artifact lies somewhere between 0.99099 and 0.999802. There is still some plausibility remaining for an epilepsy diagnosis originating from the debate, but it is very small because of the low relative reliability in comparison with the video evidence.

Note that, if we had known from the debate which proportion of the clinicians supported the epilepsy diagnosis and which proportion supported the artifact conclusion, m_1 would have distinguished the two cases as singular alternatives with the proportion as confidence (provided we assume all clinicians participating in the debate have the same reliability). Alternatively, one could also include the opinion of each clinician participating in the debate as a separate piece of evidence. The mixing rule would produce a similar combined result.

4.6.2 Second example: the hemochromatosis case

Here is how we abbreviate the names of the diseases used in this example: diabetes as diab, alcoholism as alc, hemochromatosis as hemo, hepatitis C as hepC and infection as inf.

A few days before the patient lands in the ER, a first diagnosis of diabetes type I is made. Our frame of discernment at this point contains labels diab and

other_1.

$$\begin{aligned}\mathcal{L} =& \{\mathsf{diab}, \mathsf{other}_1\}\\ w_1 =& 1\\ m_1(A) =& \begin{cases} 1 & \text{if } A = \{\mathsf{diab}\}\\ 0 & \text{if } A = \{\mathsf{other}_1\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

The patient lands in the ER and some hypotheses are first considered: hepatitis C or infection. This corresponds to a refinement of the frame of discernment to include both hepatitis C and infection, i.e., $\mathsf{other}_1 \equiv \mathsf{hepC} \sqcup \mathsf{inf} \sqcup \mathsf{other}_2$. We need to adapt m_1 to the refined frame of discernment.

$$\begin{aligned}\mathcal{L}' =& \{\mathsf{diab}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ w_1' =& 1\\ m_1'(A) =& \begin{cases} 1 & \text{if } A = \{\mathsf{diab}\}\\ 0 & \text{if } A = \{\mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

The initial consideration of hepatitis C or infection as opposed to diabetes given at the ER can be interpreted as full confidence in hepatitis C or infection. Let us suppose this interpretation is considered more reliable than the initial diabetes evidence, say twice as reliable.

$$\begin{aligned}\mathcal{L}' =& \{\mathsf{diab}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ w_2 =& 2\\ m_2(A) =& \begin{cases} 1 & \text{if } A = \{\mathsf{hepC}, \mathsf{inf}\}\\ 0 & \text{if } A = \{\mathsf{diab}, \mathsf{other}_2\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

Applying the mixing rule gives the following combined diagnosis.

$$\begin{aligned}\mathcal{L}' =& \{\mathsf{diab}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ W_{12} =& 3\\ m_{12}(A) =& \begin{cases} \frac{1}{3} & \text{if } A = \{\mathsf{diab}\}\\ 0 & \text{if } A = \{\mathsf{diab}, \mathsf{other}_2\}\\ \frac{2}{3} & \text{if } A = \{\mathsf{hepC}, \mathsf{inf}\}\\ 0 & \text{if } A = \{\mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

Both the hepatitis C and infection are quickly ruled out at the ER by means of tests and the diagnosis retained is that of diabetes combined with severe alcoholism. Ruling out the initial ER diagnosis can be achieved by updating its weight w_2 to $w_2' = 0$. The alternative "diabetes combined with severe alcoholism" is a subconcept of diabetes. Therefore, we need to apply operation 3a of Section 4.5.2 to split the label for diabetes: $\mathsf{diab} \equiv \mathsf{diab\_alc} \sqcup \mathsf{diab\_no\_alc}$. We update m_1' again (we omit m_2, because its weight is $w_2 = 0$, hence it doesn't count anymore)

$$\begin{aligned}\mathcal{L}'' =& \{\mathsf{diab\_alc}, \mathsf{diab\_no\_alc}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ w_1'' =& 1\\ m_1''(A) =& \begin{cases} 1 & \text{if } A = \{\mathsf{diab\_alc}, \mathsf{diab\_no\_alc}\}\\ 0 & \text{if } A = \{\mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

Because of the additional tests, we may consider the new evidence more reliable to a degree of $w_3 = 4$. The evidence of retaining the diagnosis of diabetes but combined with alcoholism can be represented as

$$\begin{aligned}\mathcal{L}'' =& \{\mathsf{diab\_alc}, \mathsf{diab\_no\_alc}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ w_3 =& 4\\ m_3(A) =& \begin{cases} 1 & \text{if } A = \{\mathsf{diab\_alc}\}\\ 0 & \text{if } A = \{\mathsf{diab\_no\_alc}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_2\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

However, after research by the patient and a pathologist friend, a different explanation of the symptoms comes to the scene: hemochromatosis. After several

tests, this diagnosis is confirmed. This turn of events first calls for yet another expansion of the frame of discernment: $\mathsf{other}_2 \equiv \mathsf{hemo} \sqcup \mathsf{other}_3$. Moreover, the positive testing for hemochromatosis should not be seen as a case of just some more evidence to add to the mix, but rather as evidence that overrules all previous evidence. We therefore set all weights $w_1 = w_2 = w_3 = 0$. The only remaining evidence that counts, is:

$$\begin{aligned}\mathcal{L}''' =& \{\mathsf{diab\_alc}, \mathsf{diab\_no\_alc}, \mathsf{hepC}, \mathsf{inf}, \mathsf{hemo}, \mathsf{other}_3\}\\ w_4 =& 1\end{aligned}$$

$$m_4(A) = \begin{cases} 1 & \text{if } A = \{\mathsf{hemo}\} \\ 0 & \text{if } A = \{\mathsf{diab\_alc}, \mathsf{diab\_no\_alc}, \mathsf{hepC}, \mathsf{inf}, \mathsf{other}_3\} \\ 0 & \text{if } A = \varnothing \end{cases}$$

4.6.3 Evidence combination model applied to Zadeh's counterexample

Zadeh's example (introduced in [65] and explained in Section 4.3) has become the canonical example to show that the classic Dempster-Shafer evidence combination rule is not suitable for combining highly conflicting pieces of evidence. Haenni, however, contends that the apparent counter-intuitive result of the example is due to poor modelling of the problem. While the criticism leveled by [78] may be founded, we show how our evidence combination model makes the modelling of Zadeh's example very simple and leads to a logical result.
In Zadeh's example, we have 2 sources of evidence, two clinicians giving conclusions, denoted as clinicians c_1 and c_2, and 3 alternatives (meningitis abbreviated with men, brain tumor abbreviated with tumor and concussion abbreviated with conc). The diagnosis of clinician c_1 is

$$\begin{aligned}\mathcal{L} =& \{\mathsf{men}, \mathsf{tumor}, \mathsf{conc}, \mathsf{other}\}\\ w_{c_1} =& 1\end{aligned}$$

$$m_{c_1}(A) = \begin{cases} 0.99 & \text{if } A = \{\mathsf{men}\} \\ 0.01 & \text{if } A = \{\mathsf{tumor}\} \\ 0 & \text{if } A = \{\mathsf{conc}, \mathsf{other}\} \\ 0 & \text{if } A = \varnothing \end{cases}$$

We have $m_{c_1}(\{\mathsf{conc}\}) = 0$ as there is no evidence for the concussion alternative at this point. The conclusions drawn by clinician c_2 point to a different

direction.

$$\begin{aligned}\mathcal{L} =& \{\mathsf{men}, \mathsf{tumor}, \mathsf{conc}, \mathsf{other}\}\\ w_{c_2} =& 1\\ m_{c_2}(A) =& \begin{cases} 0.99 & \text{if } A = \{\mathsf{conc}\}\\ 0.01 & \text{if } A = \{\mathsf{tumor}\}\\ 0 & \text{if } A = \{\mathsf{men}, \mathsf{other}\}\\ 0 & \text{if } A = \varnothing \end{cases}\end{aligned}$$

Applying the mixing rule gives us the following:

$$m_{12}(A) = \begin{cases} 0.495 & \text{if } A = \{\mathsf{men}\}\\ 0.01 & \text{if } A = \{\mathsf{tumor}\}\\ 0.495 & \text{if } A = \{\mathsf{conc}\}\\ 0 & \text{if } A = \{\mathsf{men}, \mathsf{other}\}\\ 0 & \text{if } A = \{\mathsf{conc}, \mathsf{other}\}\\ 0 & \text{if } A = \varnothing \end{cases}$$

The brain tumor alternative is, as expected, extremely unlikely. And since both clinicians (supposed equally reliable) give it the same likelihood, the final basic probability assignment associated with it, $m(\{\mathsf{tumor}\}) = 0.01$ is not wholly unexpected. That both the *concussion* and *meningitis* alternatives are equally likely after both clinicians' conclusions also makes sense, since at this point, there is no way to say that one alternative is more likely than the other. There is no reason to trust one clinician more than the other. Note that, in our modeling of Zadeh's example, the frame of discernment is still $\{\mathsf{men}, \mathsf{tumor}, \mathsf{conc}\}$ as in [65] and not one of the more complex frames of discernment used in [78].

4.7 Analytical validation

In this section, we analytically validate the evidence representation and combination model. We formulate and prove several correctness, monotonicity, and convergence properties. An experimental validation, i.e. a user study that shows that the model improves the decision quality in diagnostic processes, is beyond the scope of this chapter.

4.7.1 Validation of correctness properties

The mixing rule produces a mass function

To prove
The intention of the mixing rule is to combine several mass functions into a combined mass function that represents the combined evidence. Therefore, the result of the mixing rule should be a proper mass function:

$$m_{1..n}(\varnothing) = 0$$
$$\sum_{A \in \mathcal{F}} m_{1..n}(A) = 1$$

Proof
Since all m_i are mass functions, $m_i(\varnothing) = 0$ ($i \in [1..n]$). Therefore,

$$m_{1..n}(\varnothing) \;=\; \frac{1}{W}\sum_{i=1}^{n} w_i m_i(\varnothing) \;=\; \frac{1}{W}\sum_{i=1}^{n} w_i 0 \;=\; 0$$

The other requirement is that the sum of masses of all alternatives should be 1.

$$\begin{aligned}\sum_{A \in \mathcal{F}} m_{1..n}(A) \;&=\; \sum_{A \in \mathcal{F}} \frac{1}{W}\sum_{i=1}^{n} w_i m_i(A) \;=\; \frac{1}{W}\sum_{A \in \mathcal{F}}\sum_{i=1}^{n} w_i m_i(A) \\ &= \frac{1}{W}\sum_{i=1}^{n}\sum_{A \in \mathcal{F}} w_i m_i(A) \;=\; \frac{1}{W}\sum_{i=1}^{n} w_i \sum_{A \in \mathcal{F}} m_i(A) \\ &= \frac{1}{W}\sum_{i=1}^{n} w_i \cdot 1 \;=\; \frac{1}{W} W \;=\; 1\end{aligned}$$

Agreement between evidence of varying reliability

To prove
If several pieces of evidence agree with each other, their weights should not be relevant. In other words, let all mass functions m_i ($i \in [1..n]$) be equal, i.e., $\forall i \in [1..n] : m_i = m$. Then $m_{1..n}$ should be equal to m irrespective of the individual weights w_i.
Proof

$$(\forall A \in \mathcal{F}) \quad m_{1..n}(A) \;=\; \frac{1}{W}\sum_{i=1}^{n} w_i m_i(A) \;=\; \frac{1}{W}\sum_{i=1}^{n} w_i m(A)$$

$$= \frac{1}{W} m(A) \sum_{i=1}^{n} w_i \;=\; \frac{1}{W} m(A) W \;=\; m(A)$$

Therefore, $\forall i \in [1..n] : m_i = m \;\Rightarrow\; m_{1..n} = m$

Weights are relative

To prove
Weights are intended to be relative, i.e., if two pieces of evidence m_1 and m_2 are given, then it should not matter for $m_{1,2}$ whether $w_1 = 1$ and $w_2 = 3$, or that $w_1 = 5$ and $w_2 = 15$. In other words, if all weights w_i are updated with the same factor $f \neq 0$ using operation 2 (Section 4.5.2), the same $m_{1..n}$ is produced by the mixing rule.
Proof
Let $J = [1..n]$, $w'_j = f w_j$ $(j \in J)$. First observe that $W' = \sum_{j=1}^{n} w'_j = \sum_{j=1}^{n} f w_j = fW$.
According to operation 2

$$\begin{aligned}
&(\forall A \in \mathcal{F}) \\
&\quad m_{new}(A) = \frac{W}{W'} m_{old}(A) + \frac{1}{W'} \sum_{j \in J} (w'_j - w_j) m_j(A) \\
&\qquad = \frac{W}{fW} m_{old}(A) + \frac{1}{fW} \sum_{j \in J} (f w_j - w_j) m_j(A) \\
&\qquad = \frac{1}{f} m_{old}(A) + \frac{1}{fW} \sum_{j=1}^{n} w_j (f-1) m_j(A) \\
&\qquad = \frac{1}{f} m_{old}(A) + \frac{1}{f} \frac{1}{W} (f-1) \sum_{j=1}^{n} w_j m_j(A) \\
&\qquad = \frac{1}{f} m_{old}(A) + \frac{f-1}{f} m_{old}(A) \\
&\qquad = m_{old}(A) \left(\frac{1}{f} + \frac{f-1}{f} \right) \\
&\qquad = m_{old}(A) \left(\frac{1+f-1}{f} \right) \\
&\qquad = m_{old}(A)
\end{aligned}$$

The refinement of the frame of discernment does not modify existing evidence

To prove
The refinement of the frame of discernment (operations 3a and 3b of Sections 4.5.2 and 4.5.2) is a three-step procedure which updates all existing mass functions. Let us denote the refinement of a mass function with *refine* and the mixing rule with *mix*. Since it is only a refinement of a label, the mass function resulting from the mixing rule should be equal to a direct refinement of the combined evidence. In other words

$$\textit{refine}(m_{1..n}) = \textit{mix}(\textit{refine}(m_1), \ldots, \textit{refine}(m_n))$$

Proof
refine(m_i) produces a m'_i in the following way:

$$(\forall A' \in \mathcal{F}') \quad \begin{cases} m'(A') = m(A) & \text{if } \exists A \in \text{dom}(m) : A' = \textit{refine}(A) \\ 0 & \text{otherwise} \end{cases}$$

where $(\forall A \in \mathcal{F})$ *refine*$(A) = \begin{cases} (A \backslash \{a\}) \cup \{a_1, \ldots, a_m\} & \text{if } a \in A \\ A & \text{otherwise} \end{cases}$

Note that $\forall A' \in \mathcal{F}'$, this A' either has a corresponding $A \in \mathcal{F}$ or not. In the former case, the corresponding A is the inverse refinement, i.e., $A = (A' \backslash \{a_1, \ldots, a_m\}) \cup \{a\}$. In this case, $\forall A \in \mathcal{F}\, \forall i \in [1..n] : m_i(A) = m'_i(A')$. In the second case where $\exists a_j \in A'$ but $\{a_1, \ldots, a_m\} \not\subseteq A'$, $m'_i(A') = 0$.
Applying the mixing rule

$$\begin{aligned} & (\forall A' \in \mathcal{F}') \\ m'_{1..n}(A') &= \frac{1}{W} \sum_{i=1}^{n} w_i \textit{refine}(m_i)(A') \\ &= \begin{cases} \frac{1}{W} \sum_{i=1}^{n} w_i m_i(A) & \text{if } \exists A \in \text{dom}(m_i) : A' = \textit{refine}(A) \\ \frac{1}{W} \sum_{i=1}^{n} w_i \cdot 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} m_{1..n}(A) & \text{if } \exists A \in \text{dom}(m_i) : A' = \textit{refine}(A) \\ 0 & \text{otherwise} \end{cases} \\ &= \textit{refine}(m_{1..n})(A') \end{aligned}$$

4.7.2 Validation of monotonicity properties

Support for an alternative should increase its confidence

To prove
If new evidence (operation 1 from Section 4.5.2) positively supports an alternative A, i.e, $m_u(A) > m_{old}(A)$, then the confidence in A should increase irrespective of the weight w_u, i.e., $m_{new}(A) > m_{old}(A)$.
Proof
Since $W' = W + w_u$ and weights are positive, we know that $0 < \frac{W}{W'} < 1$ and $0 < \frac{w_u}{W'} < 1$.

$$\begin{aligned} m_{new}(A) &= \frac{W}{W'} m_{old}(A) + \frac{w_u}{W'} m_u(A) \\ &> \frac{W}{W'} m_{old}(A) + \frac{w_u}{W'} m_{old}(A) \\ &= m_{old}(A) \left(\frac{W + w_u}{W'} \right) \\ &= m_{old}(A) \end{aligned}$$

Increase in weight for evidence supporting an alternative should increase its confidence

To prove
If a weight is increased (operation 2 from Section 4.5.2) for evidence that positively supports an alternative A, i.e, $w'_j > w_j$ for some $j \in [1..n]$ and $m_j(A) > m_{old}(A)$, then the confidence in A should increase, i.e., $m_{new}(A) > m_{old}(A)$. Let $J = \{j\}$.
Proof

$$\begin{aligned} m_{new}(A) &= \frac{W}{W'} m_{old}(A) + \frac{1}{W'} \sum_{j \in J} (w'_j - w_j) m_j(A) \\ &= \frac{W}{W'} m_{old}(A) + \frac{(w'_j - w_j) m_j(A)}{W'} \\ &> \frac{W}{W'} m_{old}(A) \quad \text{because } w'_j > w_j \wedge m_j(A) > 0 \wedge W' > 0 \\ &> m_{old}(A) \quad \text{because } W' > W \end{aligned}$$

4.7.3 Validation of convergence properties

Continuous positive or negative evidence cancels out initial uncertainty

To prove
Suppose we have an initial uncertain evidence, i.e., $0 < m_1(A) < 1$ for some alternative A (hence also for one or more other alternatives in $\text{dom}(m_1)$). If we were to give continuous positive evidence on that alternative, i.e., $\forall i > 1 : m_i(A) = 1$, then the initial uncertainty for that alternative cancels out, i.e., $\lim_{n\to\infty} m_{1..n}(A) = 1$. Analogously, continuous negative evidence on that alternative will also cancel out the uncertainty but towards 0, i.e., if $\forall i > 1 : m_i(A) = 0$, then $\lim_{n\to\infty} m_{1..n}(A) = 0$. Let $W_n = \sum_{i=1}^{n} w_i$.

Proof
According to operation 1 (Section 4.5.2)

$$
\begin{aligned}
(\forall A \in \mathcal{F}) \quad \lim_{n\to\infty} m_{1..n}(A) &= \lim_{n\to\infty} \frac{1}{W_n} \sum_{i=1}^{n} w_i m_i(A) \\
&= \lim_{n\to\infty} \frac{1}{W_n} \left(w_1 m_1(A) + \sum_{i=2}^{n} w_i m_i(A) \right) \\
&= \lim_{n\to\infty} \frac{1}{W_n} \left(w_1 m_1(A) + \sum_{i=2}^{n} w_i \right) \\
&= \lim_{n\to\infty} \frac{1}{W_n} \left(w_1 m_1(A) + (W_n - w_1) \right) \\
&= \lim_{n\to\infty} \left(\frac{w_1 m_1(A)}{W_n} + \frac{W_n - w_1}{W_n} \right) \\
&= 0 + 1 = 1
\end{aligned}
$$

Because, in the limit of $n \to \infty$, the term $w_1 m_1(A)$ remains constant, while $\lim_{n\to\infty} W_n = \infty$ and $\lim_{n\to\infty} \frac{W_n - w_1}{W_n} = 1$.
For the continuous negative evidence case, we get

$$
\begin{aligned}
(\forall A \in \mathcal{F}) \quad \lim_{n\to\infty} m_{1..n}(A) &= \lim_{n\to\infty} \frac{1}{W_n} \sum_{i=1}^{n} w_i m_i(A) \\
&= \lim_{n\to\infty} \frac{1}{W_n} \left(w_1 m_1(A) + \sum_{i=2}^{n} w_i m_i(A) \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \frac{1}{W_n} \left(w_1 m_1(A) + 0\right) \\
&= \lim_{n\to\infty} \frac{w_1 m_1(A)}{W_n} \\
&= 0
\end{aligned}
$$

Because in the limit of $n \to \infty$ the term $w_1 m_1(A)$ remains constant, while $\lim_{n\to\infty} W_n = \infty$.

Random pick

To prove
Suppose we have some frame of discernment $\mathcal{F}$ and we continuously give random evidence with the same weight, then we expect the combined evidence to even out, i.e., evaluate each possible diagnosis as equally likely. We focus on the following form of giving random evidence. We pick a label a randomly and provide new evidence

$$
m(A) = \begin{cases} 1 & \text{if } A = \{a\} \\ 0 & \text{otherwise} \end{cases}
$$

Proof
Note that we only give evidence on singular alternatives $\{a\}$. Therefore, $m_i(A) = 0$ for non-singular alternatives, hence $m_{1..n}(A) = 0$ as well. Let k be the number of singular alternatives, i.e., $k = |\mathcal{L}|$. Let $J_a = \{i \in [1..n] \mid m_i(\{a\}) = 1\}$. According to operation 1 (Section 4.5.2)

$$
\begin{aligned}
(\forall a \in \mathcal{L}) \quad \lim_{n\to\infty} m_{1..n}(\{a\}) &= \lim_{n\to\infty} \frac{1}{W_n} \sum_{i=1}^{n} w_i m_i(\{a\}) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i\in J_a} m_i(\{a\}) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i\in J_a} 1 \\
&= \lim_{n\to\infty} \frac{|J_a|}{n} \\
&= \frac{1}{k}
\end{aligned}
$$

because we pick a randomly out of k possibilities, hence $|J_a| = \frac{n}{k}$ when $n \rightarrow \infty$.

4.8 Storing evidence with lineage in a probabilistic database

4.8.1 Probabilistic databases

There are several representation systems for uncertain data. The ULDB model (Uncertainty-Lineage Databases) is an extension of the classic RDBMS model where data uncertainty and lineage are stored alongside the data itself [79, 80, 81]. This model, as well as several others in the database community such as MayBMS [82, 83], is based on possible worlds semantics. An important application domain for probabilistic databases is data integration [84], because the explicit treatment of the uncertainty surrounding (integrated) data is an important technique for achieving better data quality [85].

While the ULDB-model has been shown to be adequate for storing aleatory uncertainty, its quantification of uncertainty in terms of probabilities doesn't allow it to store data with epistemic uncertainty. Examples of epistemic uncertainty are missing confidence values and imprecisely or coarsely defined confidence values. [86] introduces a generalization of the ULDB-model in which confidence values are taken to be basic probability assignments as introduced by the Dempster-Shafer theory. The generalized model still follows the possible worlds semantics. Since in practice, medical data, in particular medical diagnosis, is rarely defined in terms of precise probabilities, we believe that the generalized ULDB model introduced in [86] is suitable for storing evidence data in compliance with the model described in this chapter.

Being able to incorporate meta-evidence, i.e., evidence that says something about earlier evidence, such as invalidating it or reducing its reliability, is useful and reflects the true evolution of a diagnosis during the diagnosis determination process. The importance of meta-evidence has been confirmed in other domains. Several feedback models have been proposed. [75, 76] use Bayesian networks to implement relevance feedback. [87] and [88] show how user feedback (positive and negative feedback) can be used to improve the quality of data within their proposed probabilistic XML data integration framework (based on the possible worlds semantics). It is important to observe that these models are based on classic probability theory and are restricted to posi-

tive and negative feedback. Feedback that introduces a new alternative, which is a frequent type of feedback as shown in Section 4.5, is not supported. Moreover, epistemic uncertainty cannot be properly represented. Therefore, we consider that such models, although supporting meta-evidence in the form of feedback, are unsuitable for our purpose of supporting the incremental diagnostis process.

4.8.2 Data representation

In general, a DBMS is a robust and safe place for storing, querying and manipulating important data. Medical data, in particular the EEGs in this chapter, as well as the pieces of evidence collected during the diagnosis process would benefit from residing in a database. As argued above, probabilistic database technology is well suited to store medical data and associated evidence because of its native handling of uncertainty.

We have chosen the generalized ULDB-model as described in [86] as a database model. While the generalized uncertain database model preserves the concepts of *alternatives*, *x-tuples* and *lineage*, it substitutes the probabilities assigned to each alternative with a basic probability assignment. We refer to [86] for the representation of the other cases (coarse confidence values and imprecise confidence values). In everything that follows, we will not consider 'maybe' annotations although they can in theory exist in a generalized uncertain database model. Note that our model is not really specific to a relational database and may be used with few alterations with other types of databases, most notably XML databases [88].

Appendix A provides a description of a proof of concept implementation of EEG data and evidence storage system. In short, each piece of evidence is stored as an x-tuple with an ULDB-alternative associated to each alternative A. The x-tuple also contains attributes that refer to associated data such as the EEG-files. Note that the mixing rule as well as the atomic operations are in the worst case linear in the number of pieces of evidence and alternatives and the number of x-tuples (evidences) is small in comparison with ordinary business data. Furthermore, we have shown that some of the atomic operations can be computed incrementally reducing the complexity even further. Therefore, no scalability problems are to be expected.

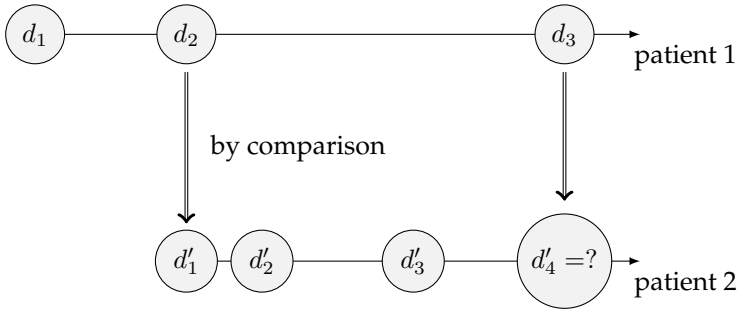


Figure 4.6: Motivation

4.8.3 Lineage and versioning

Another feature commonly supported in probabilistic databases is *lineage* or *data provenance*. By *lineage* or *data provenance*, we mean meta-data on which piece of data was derived from which other pieces of data.

Figure 4.6 illustrates the importance of lineage for a medical diagnosis support system. Suppose a clinician has two patients: patient 1 and patient 2. Patient 1 is initially diagnosed at d_1 and subsequently his diagnosis was revised in d_2. At this point, the clinician seeking to diagnose patient 2 notices similarities between patient 1 and patient 2's cases and reaches a first diagnosis (d'_1) for patient 2 by comparison with patient 1's current diagnosis (d_2). Patient 2's diagnosis was subsequently revised (d'_2 and d'_3). At one point in time, new evidence leads to a new diagnosis d_3 for patient 1. Given that patient 2's diagnosis was derived from patient 1's diagnosis, it is important to trace back these derivations for determining how the new evidence leading to d_3 affects patient 2's diagnosis. Lineage is needed in the database to be able to do this retracing.

Furthermore, for auditing and quality assurance purposes as well as for rolling back wrong modifications (e.g., faulty evidence due to material problems distorting test results or fraudulent behavior by a clinician), it is important to be able to review all modifications to pieces of evidence. Operations 2, 3a, and 3b all modify the evidence data in the databases. Standard versioning support is sufficient to allow the retracing not only of derivation chains, but also of all modifications that happened during a diagnostic process.

4.9 Conclusion

In this chapter, we propose an evidence combination model targeted at medical diagnostic processes that (a) is based on Dempster-Shafer theory, (b) provides a combined diagnosis derived from all known evidence and opinions known at a certain point in time, (c) allows the refinement of considered alternatives as well as the addition of new alternatives during the diagnosis process, and (d) supports the inclusion of meta-evidence, i.e., feedback from one clinician on the diagnosis of another or evidence that reduces or nullifies the reliability of an earlier diagnosis. Furthermore, it can handle situations where confidence values are either missing (by far the most frequent case in practice), or coarsely or imprecisely defined. In other words, it tackles both aleatory and epistemic uncertainty.

The model itself is based on evidence being represented by mass functions, a mixing rule based on a normalized weighted average, and 3 atomic operations: adding new evidence, updating weights, and refining the considered alternatives by either splitting a label or adding new alternatives. The model has been validated analytically by proving several correctness, monotonicity, and convergence properties. We also discussed how the model can be implemented with probabilistic database technology. We have illustrated the model with two running examples: 1. the toothbrush case where an initial disagreement between clinicians was settled by a later viewing of a video recording which proved that the abnormal EEG segment was the result of the patient brushing his teeth. 2. the hemochromatosis case where a repeated diagnosis for diabetes/alcoholism which didn't even really match all symptoms was overturned late in the process by a diagnosis of a rare condition called hemochromatosis not considered initially.

In the introduction, we pointed out the fact that the evidence combination model was a first step into the process of building a system to support the medical diagnosis process. The goal of the diagnosis support system is to increase the chances of an early correct diagnosis. An important building block for such a system is an automatic component (for instance, a rule-based, similarity search-based or machine learning-based component) that, given the often disregarded patient history in his/her dossier and all the clues at hand, outputs a list of likely alternatives longer than usually considered by clinicians for efficiency reasons— be the alternatives rare conditions ("zebras") or common ones. Rare conditions may enter the scene if multiple clues pointing to a rare condition and unlikely to occur together without said rare condition are found.

Such an automatic diagnosis can be mixed with the other pieces of evidence using the model in this chapter. In this way, it can bring history-based evidence and rare conditions worthy of consideration to the attention of the clinician in an unobtrusive way early in the process, so that correct diagnoses and treatments are deduced faster in complex and ambiguous situations. It would be at best less likely to overlook "zebras" and in the worse case the "zebras" would overlooked for a shorter period of time.

Acknowledgments

We would like to thank Michel van Putten from the Technical Medicine department at the University of Twente for providing the EEG data shown in this paper as well as the data and specifics linked to the toothbrush case example used in this paper.

CHAPTER 5

Similarity search on EEG data

Part of this chapter was published as [52].

Chapter 3 introduced a platform that can be used to share medical data and process it in reasonable times. This chapter proposes a feature-based similarity measure for similarity search on EEG data. Candidate features are the fractal dimension, spectral entropy and high/low frequency ratio.

5.1 Motivation

In Chapter 3, we proposed a Hadoop-based platform for medical data sharing. This would, in theory, make a vast trove of patient data accessible for clinicians in their diagnosis-forming process as well as to researchers seeking to develop medical data automated interpretation methods.

With this trove of data, researchers may seek to construct balanced annotated datasets to develop and test automated medical data interpretation methods. And clinicians may rely on the shared patient data to perform diagnosis by comparison (eg. based on abnormal/eventful EEG subsegment). Browsing the data in the shared medical repository through similarity search would allow researchers to construct the datasets they need and clinicians to perform diagnoses by comparison.

Similarity search approaches would vary depending on the nature of the data searched, eg. images or time series. This chapter only focuses on EEG data i.e multidimensional time series type of data.

Similarity search is a complement to queries based on existing annotations. Similarity search can be performed by defining a similarity metric between EEG segments followed by a ranking of the EEG segments (from the data

repository) according to their similarity to the EEG segment that constitutes the query/request. The goal of the similarity search approach can be summed up as follows: given a query EEG segment Q, EEG segments that are similar to Q or EEGs that contain at least an occurrence of Q can be retrieved from the EEG data repository even when the EEG repository is not fully annotated (which is often the case since annotating a complete data repository is time-consuming and costly).

EEG recordings correspond to very diverse conditions (eg. "normal" state, seizure episodes, Alzheimer disease). Trying to use disease-specific features and similarity metrics would require an exponential number of features and similarity metrics to be defined. Therefore, a generic similarity metric, that relies on the fact that EEGs are multidimensional time series, is required. The goal of similarity search is not after all to perform automatic EEG classification and interpretation. Similarity search on EEG data is more comparable to a rougher "Google-"like data retrieval approach based on queries that are multidimensional time series segments rather than plain text queries. Note that, in this context, similarity search is already considered successful if it results in retrieving only a few relevant segments at the top of the list of segments (ranked by similarity to the query, the most similar appearing on top) returned by the similarity search. In fact, even if only a few relevant segments are returned in the top 10 ranking, their being in the top 10 ranking guarantees that they are easily accessible to the user, in particular if at the top of the list, and therefore that relevant information is easily accessible to the user.

In Section 5.3, we explain why canonical time series similarity methods that on modeling the time series with an ARIMA/AR model followed by feature extraction based on the model may not be suitable for EEG data and then study the discriminative performance of a simple similarity measure based on the fractal dimension. Some background on the fractal dimension is given in Section 5.2. Then in Section 5.4, we study EEG similarity search performed with a feature-based based similarity measure (the features tested being the fractal dimension, the spectral entropy and the high/low frequency ratio).

5.2 Some background on fractal interpolation and fractal dimension

The similarity measure we propose in Section 5.3 relies on the fractal dimension. Therefore, in this section, we give a definition of the fractal dimension

(5.2.1) and explain some of its applications.
The fractal dimension can be computed using various methods ([89, 90]). In this chapter, we use two methods, one method for Section 5.3 and another for Section 5.4.
The method we use in Section 5.3 consists in interpolating the time series segment with a set of functions called iterated function systems (IFS) and estimating the fractal dimension based on the IFS interpolation parameters. We give some background on each component of this method and in particular give a definition of IFS (Section 5.2.2), explain how they can be used to interpolate time series segments (Section 5.2.3) and give the theorem that links the IFS interpolation parameters of a times series segment to its fractal dimension (Section 5.2.4).
The second method, used to calculate the fractal dimension in Section 5.4, is the Petrosian method. This method is described in 5.2.5.

5.2.1 Fractal dimension

Every object can be defined thanks to a property called topological dimension or Lebesgue Covering dimension. A space/object has a Lebesgue Covering dimension n if, for every open cover [1] of that space, there is an open cover that refines it such that the refinement [2] has order at most $n + 1$. For example, the topological dimension of the Euclidean space $\mathbb{R}^n$ is n, the topological dimension of a point is 0 and the topological dimension of a curve (eg line or time series) is 1. The topological dimension, however, measures the local size of a space in a very crude way: for instance, both a straight line and a curve that almost fills a plane have a topological dimension of 1.
The fractal dimension, sometimes called Hausdorff dimension, is an extension of the topological dimension that can be seen as the statistical quantity that gives an indication of how completely an object appears to fill space, as one zooms down to finer and finer scales. The fractal dimension counts the effective number of degrees of freedom of an object and therefore quantifies its complexity. The fractal dimension is equal to the topological dimension for

[1] A covering of a subset S is a collection $\mathcal{C}$ of open subsets in X whose union contains all of S at least. A subset $\mathcal{S} \subset X$ is open if it is an arbitrary union of open balls in X. This means that every point in $\mathcal{S}$ is surrounded by an open ball which is entirely contained in X. An open ball in a in metric space X is defined as a subset of X of the form $B(x_0, \epsilon) = \{x \in X | d(x, x_0) < \epsilon\}$ where x_0 is a point of X and ϵ a radius.

[2] A refinement of a covering $\mathcal{C}$ of S is another covering $\mathcal{C}'$ of S such that each set B in $\mathcal{C}'$ is contained in some set A in $\mathcal{C}$

smooth shapes or shapes with few corners, for example straight lines, planes, cubes. In other cases, the fractal dimension is generally a non-integer or fractional number. Typically, for a time series, the fractal dimension is comprised between 1 and 2 since the (topological) dimension of a plane is 2 and that of a line is 1.
The fractal dimension has been used to:

- uncover patterns in datasets and cluster data ([91, 92, 93])
- analyse medical time series ([94, 95]) such as EEGs ([96, 14])
- determine the number of features to be selected from a dataset for a similarity search while obviating the "dimensionality curse" ([97])

There are several ways to calculate the fractal dimension of a time series segment. One way (used in Section 5.3) consists in interpolating said time series segment with an iterated function system (IFS) and then estimating the fractal dimension based on the IFS parameters found through interpolation and Barnsley's theorem linking those parameters and the fractal dimension. Section 5.2.2 gives a definition of IFS while Section 5.2.3 explains how to perform fractal interpolation and Section 5.2.4 gives the theorem that links fractal interpolation parameters with the fractal dimension. Another way to compute the fractal dimension of a time series is the Petrosian method described in [90] and used in Section 5.4. We give a brief explanation of the Petrosian method in 5.2.5. [89] describes and compares other methods used to compute the fractal dimension (in addition to the Petrosian method).

5.2.2 Iterated function systems

We denote as $\mathbf{K}$ a compact metric space for which a distance function d is defined and as $\mathbb{C}(\mathbf{K})$ the space of continuous functions on $\mathbf{K}$. We define over $\mathbf{K}$ a finite collection of mappings $\mathbb{W} = w_{ii\in[1,n]}$ and their associated probabilities $p_{ii\in[1,n]}$ such that :

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} p_i = 1$$

.
We also define an operator T on $\mathbb{C}(\mathbf{K})$ as $(Tf)(x) = \sum_{i=1}^{n} p_i(f \circ w_i)(x)$. If T maps $\mathbb{C}(\mathbf{K})$ into itself, then the pair (w_i, pi) is called an iterated function system

on $(\mathbf{K}, d)$. The condition on T is satisfied for any set of probabilities p_i if the transformations w_i are contracting, in other words, if, for any i, there exists a $\delta_i < 1$ such that: $d(w_i(x), w_i(y)) \leq \delta_i d(x, y) \quad \forall x, y \in K$. The IFS is also denoted as hyperbolic in this case.

5.2.3 Principle of fractal interpolation

If we define a set of points $(x_i, F_i) \in \mathbb{R}^2 : i = 0, 1, ..., n$ with $x_0 < x_1 < ... < x_n$, then an interpolation function corresponding to this set of points is a continuous function $f : [x_0, x_n] \to \mathbb{R}$ such that $f(x_i) = F_i \quad$ for $i \in [0, n]$.
In fractal interpolation, the interpolation function is often constructed with n affine maps of the form:

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} \quad i = 1, 2, ..., n$$

where d_i is constrained to satisfy: $-1 \leq d_i \leq 1$. Furthermore, we have the following constraints:

$$w_i \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix} \quad \text{and} \quad w_i \begin{pmatrix} x_n \\ y_n \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

After determining the contraction parameter d_i, we can estimate the four remaining parameters (namely a_i,c_i,e_i,f_i):

$$a_i = \frac{x_i - x_{i-1}}{x_n - x_0} \tag{5.1}$$

$$c_i = \frac{x_n x_{i-1} - x_0 x_i}{x_n - x_0} \tag{5.2}$$

$$e_i = \frac{y_i - y_{i-1}}{x_n - x_0} - d_i \frac{y_n - y_0}{x_n - x_0} \tag{5.3}$$

$$f_i = \frac{x_n y_{i-1} - x_0 y_i}{x_n - x_0} - d_i \frac{x_n y_0 - x_0 y_n}{x_n - x_0} \tag{5.4}$$

d_i can be determined using the geometrical approach given in [98]. Let t be a time-series with end-points (x_0, y_0) and (x_n, y_n), and (x_p, y_p) and (x_q, y_q) two

consecutive interpolation points so that the map parameters desired are those defined for w_p. We also define α as the maximum height of the entire function measured from the line connecting the end-points (x_0, y_0) and (x_n, y_n) and β as the maximum height of the curve measured from the line connecting (x_p, y_p) and (x_q, y_q). α and β is positive (respectively negative) if the maximum value is reached above the line (respectively below the line). The contraction factor d_p is then defined as $\frac{\beta}{\alpha}$. This procedure is also valid when the contraction factor is computed for an interval instead of for the whole function. The end-points are then taken as being the end-points of the interval.

For more details on fractal interpolation, see [99, 98].

5.2.4 Estimation of the fractal dimension from a fractal interpolation

The theorem that links the fractal interpolation function and its fractal dimension is given in [99].The theorem is as follows:

Theorem 5.2.1. *Let n be a positive integer greater than 1, $\{(x_i, F_i) \in \mathbb{R}^2 : i = 1, 2, ..., n\}$ a set of points and $\{\mathbb{R}^2; w_i, i = 1, 2, .., n\}$ an IFS associated with the set of points where:*

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_i & 0 \\ c_i & d_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e_i \\ f_i \end{pmatrix} \quad \textit{for} \quad i = 1, 2, .., n.$$

The vertical scaling factors d_i satisfy $0 \leq d_i < 1$ and the constants a_i,c_i,e_i and f_i are defined as in section 5.2.3 (in equations 5.1,5.2,5.3 and 5.4) for $i = 1, 2, ..., n$. We denote G the attractor of the IFS such that G is the graph of a fractal interpolation function associated with the set of points.

If $\sum_{i=1}^{n} |d_i| > 1$ and the interpolation points do not lie on a straight line, then the fractal dimension of G is the unique real solution D of

$$\sum_{i=1} |d_i| a_i^{D-1} = 1 \tag{5.5}$$

.

5.2.5 Petrosian fractal dimension

In Sections 5.2.2, 5.2.3 and 5.2.4, we explained how to apply fractal interpolation on a time series segment and how the fractal interpolation parameters are linked to the fractal dimension through the Barnsley theorem, making it possible to estimate a segment's fractal dimension given its fractal dimension parameters. This section describes another method used to compute the fractal dimension of a time series segment: the Petrosian method.

Given a time series, the Petrosian fractal dimension (PFD) is computed based on the time series length (N) and the number of signal changes (N_δ) of the time series signal first order derivative (δ) with the following equation:

$$\frac{\log_{10} N}{\log_{10} N + \log_{10}(N/(N + 0.4N_\delta))} \tag{5.6}$$

The Petrosian fractal dimension was initially designed as a way to compute a reliable estimate of the fractal dimension of an EEG segment so as to distinguish pre-ictal segments (i.e segment preceding an epileptic seizure) and be able to predict the advent of epileptic seizures [100].

5.3 A fractal dimension-based similarity measure

5.3.1 Motivation

As explained earlier in the chapter (Section 5.1), we need to define a similarity measure to perform similarity search on EEG data. Some of the similarity measures proposed include a function interpolation step, be it piecewise linear interpolation or interpolation with AR (as in [11] to distinguish between normal EEGs and EEGs originating from the injured brain undergoing transient global ischemia) or ARIMA models, that can be followed by a feature extraction step (eg. computation of LPC cepstral coefficients from the ARIMA model of the time series as in [101]). However, ARIMA/AR methods assume that the EEG signal is stationary, which is not a valid assumption. In fact, EEG signals can only be considered as stationary during short intervals, especially intervals of normal background activity, but the stationarity assumption does not hold during episodes of physical or mental activity, for example during changes in alertness and wakefulness, during eye blinking and during transitions between various ictal states. Therefore, EEG signals are quasi-stationary.

In view of that, we propose an EEG similarity measure that is based on IFS[3] interpolation since fractal interpolation does not assume stationarity of the data and can adequately model complex structures. Furthermore, using fractal interpolation makes computing features such as the fractal dimension simple (see theorem 5.2.1 for the link between fractal interpolation parameters and fractal dimension) and the fractal dimension of EEGs is known to be a relevant marker for some pathologies such as dementia (see [14]).

5.3.2 Summary of the similarity measure computation and evaluation approach

The proposed algorithm contains 5 steps, summarized in Figure 5.1:

1. Divide each EEG channel into fixed length windows. Each EEG then contains N_i windows with $i = 1, ..M$ (M being the number of EEGs).

2. Apply fractal interpolation on each EEG window on a channel per channel basis. The result is a matrix of size $n_{ch} \times 5N_i$ where n_{ch} is the number of initial EEG channels and N_i is the number of fixed length windows per EEG. Note that n_{ch} is usually equal to 19 for EEGs recorded in the International 10/20 System. See Section 5.3.4 for details.

3. Estimate the fractal dimension of each EEG window on a channel per channel basis based on its fractal interpolation parameters. The result is a matrix of fractal dimensions of size $n_{ch} \times N_i$. See Section 5.3.4 for details.

4. With each EEG represented by the matrix obtained in step and given a similarity measure defined in 5.3.4, compute a similarity matrix between EEGs

5. Cluster the EEG with the k-medoid algorithm (described in Section 5.3.3) based on the similarity matrix obtained

[3] iterated function systems

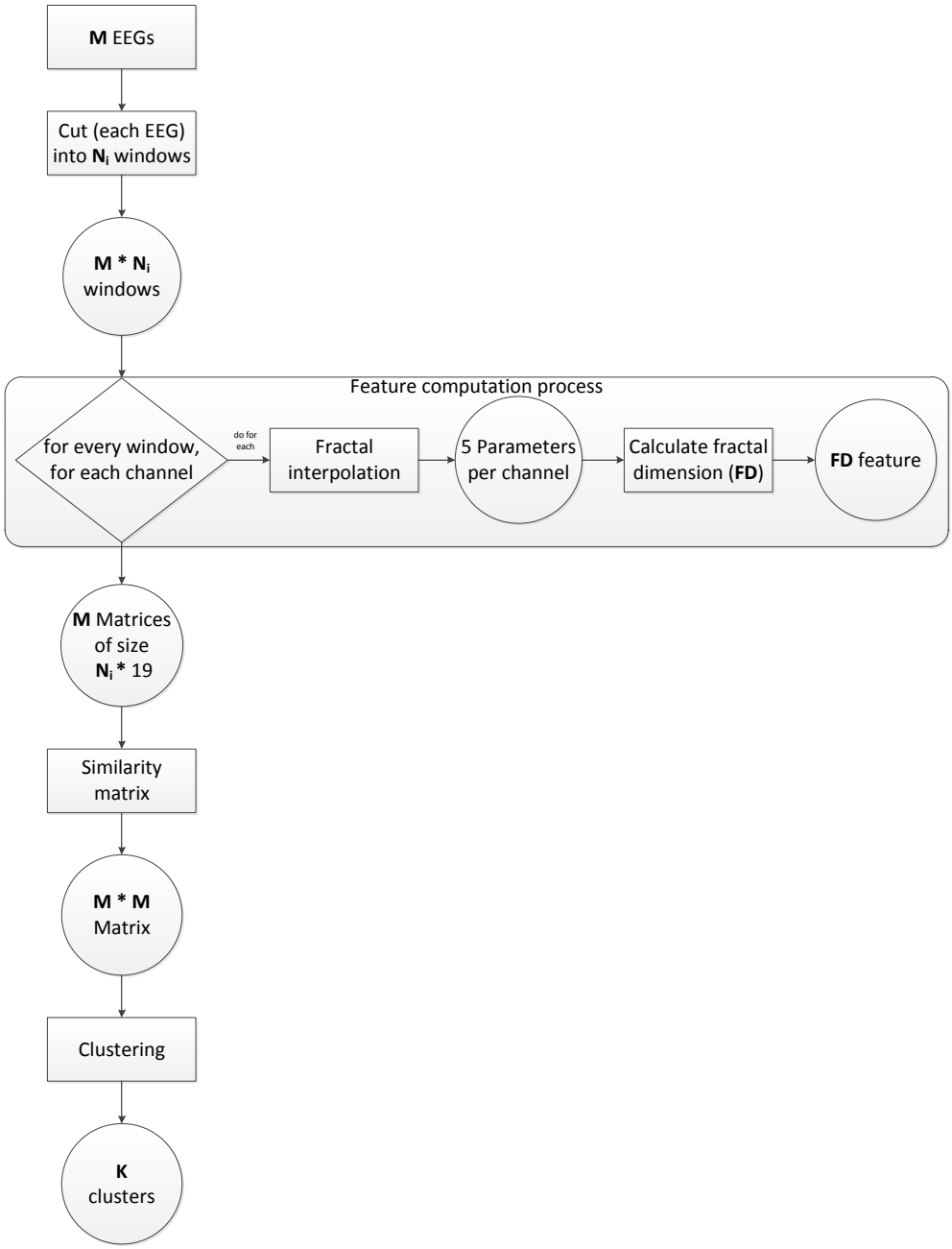


Figure 5.1: Summary of the similarity measure computation and evaluation approach for EEGs recorded in the International 10/20 System (therefore with $n_{ch} = 19$)

5.3.3 Some background on K-medoid clustering

Principle of K-medoid clustering

An $M \times M$ symmetric similarity matrix S can be associated to the EEGs to be compared (with M being the number of EEGs to compare):

$$S = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1M} \\ d_{12} & d_{22} & \dots & d_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{1M} & d_{2M} & \dots & d_{MM} \end{pmatrix} \tag{5.7}$$

where d_{nM} is the distance between EEGs n and M

Given the computed similarity matrix S (defined by equation 5.7), we can use the k-medoids algorithm to cluster the EEGs. This algorithm requires the number of clusters k to be known. We describe our choice of the number of clusters below, in the next paragraph.
The k-medoids algorithm is similar to k-means and can be applied through the use of the EM[4] algorithm. k random elements are, initially, chosen as representatives of the k clusters. At each iteration, a representative element of a cluster is replaced by a randomly chosen nonrepresentative element of the cluster if the selected criterion (e.g. mean-squared error) is improved by this choice. The data points are then reassigned to their closest cluster, given the new cluster representative elements. The iterations are stopped when no reassignments is possible. We use the PyCluster function `kmedoids` described in [108] to make our k-medoids clustering.

Choice of number of clusters

The number of clusters in the dataset is estimated based on the similarity matrix obtained following the steps in section 5.3.4 and using the method described in [109]. The method described in [109] takes the similarity matrix and

[4]short for expectation maximization algorithm. The EM algorithm is an iterative algorithm which estimates the maximum likelihood a posteriori (MAP) of the parameters of a model that depends on unobserved latent variables. It is an algorithm that alternates between two steps: an expectation (E) step that creates an expectation of the log-likelihood function using the current model parameter estimates and a maximization (M) step which computes the parameters that maximize the expected log-likelihood found in the E step. For more details on the algorithm, see [102, 103, 104, 105, 106, 107]

outputs a vector called envelope intensity associated to the similarity matrix. The number of distinct regions in the plot of the envelope intensity versus the index gives an estimation of the number of clusters. For details on how the envelope intensity vector is computed, see [109].

5.3.4 An IFS-based similarity measure

Fractal interpolation step

We interpolate each channel of each EEG (except the annotations channel) using piecewise fractal interpolation. For this purpose, we split each EEG channel into windows and then estimate the IFS for each window.
This description implies that a window size has to be chosen before estimating the piecewise fractal interpolation function for each channel.

For a relevant window size to be chosen, two parameters have to be determined: an embedding dimension and a lag τ. The size of the window is then defined as the product between lag and embedding dimension. The embedding dimension is determined thanks to Takens' theorem which states that, for the attractor of a time series to be reconstructed correctly (i.e the same information content is found in the state (latent) and observation spaces), the embedding dimension denoted m satisfies : $m > 2D + 1$ where D is the dimension of the attractor, in other words its fractal dimension. Since the fractal dimension of a time series is between 1 and 2, we can get a satisfactory embedding dimension as long as $m > 2 * 2 + 1$ i.e $m > 5$. We therefore choose an embedding dimension equal to 6. And we choose the lag τ between different elements of the delay vector to be equal to the average duration of an EEG data record i.e 1s. Therefore, we split our EEGs in (non-overlapping) windows of 6 seconds. A standard 20-minutes EEG (which therefore contains about 1200 data records of 1 second) would then be split in about 200 windows of 6 seconds.

Each window is subdivided into intervals of one second each and the endpoints of these intervals are taken as interpolation points. This means there are 7 interpolation points per interval: the starting point p_0 of the window, the point one second away from p_0, the point two seconds from p_0, the point three seconds away from p_0, the point four seconds away from p_0, the point five seconds away from p_0 and the last point of the window. The algorithm[5] to compute the fractal interpolation function per window is as follows:

[5] inspired from [98]

1. Choose, as an initial point, the starting point of the interval considered (the first interval considered is the interval corresponding to the first second of the window).
2. Choose, as the end point of the interval considered, the next interpolation point.
3. Compute the contraction factor d for the interval considered.
4. If $|d| > 1$ go to 2, otherwise go to 5.
5. Form the map w_i associated with the interval considered. In other words, compute the a, c, e and f parameters associated to the interval (see equations). Apply the map to the entire window (i.e six seconds window) to yield $w_i \begin{pmatrix} x \\ y \end{pmatrix}$ for all x in the window.
6. Compute and store the distance between the original values of the time series on the interval considered (i.e the interval constructed in steps 2 and 3) and the values given by w_i on that interval. A possible distance is the Euclidean distance.
7. Go to 2 until the end of the window is reached.
8. Store the interpolation points and contraction factor which yield the minimum distance between the original values on the interval and the values yielded by the computed map under the influence of each individual map in steps 5 and 6.
9. Repeat steps from 1 to 8 for each window of the EEG channel.
10. Apply steps 1 to 9 to all EEG channels.

Fractal dimensions estimation

After this fractal interpolation step, each window of each signal is represented by 5 parameters instead of by `signal frequency.window duration` points. The dimension of the analysed time series is therefore reduced in this step. For a standard 20-minutes EEG containing 23 signals of frequency 250 Hz, this amounts to representing each signal with 1000 values instead 50000 and the whole EEG with 23000 values instead of 1150000, and thus to reducing the

number of signal values by almost 98%. This dimension reduction may be exploited in future work to compress EEGs and store compressed representations of EEGs in the database instead of raw EEGs as the whole EEGs can be reconstructed from their fractal interpolations. Further work needs to be done on the compression of EEG data using fractal interpolation and the loss of information that may result from this compression.

Then, for each EEG channel and for each window, we compute the fractal dimension thanks to theorem 5.2.1. Equation 5.5 (of theorem 5.2.1) is solved heuristically for each 6-second interval of each EEG signal using a bisection algorithm. As we know that the fractal dimension for a time series is between 1 and 2, we search a root of the equation of theorem 5.2.1 in the interval [1,2] and split the search interval by half at each iteration until the value of the root is approached by an ϵ-margin (ϵ being the admissible error on the desired root, we choose $\epsilon = 0.0001$ in our experiments). Therefore, for each EEG channel, we have the same number of computed fractal dimensions as the number of windows. This feature extraction extraction step (fractal dimension computations) further reduces the dimensionality of the analysed time series. In fact, the number of values representing the time series is divided by 5 in this step. This leads to representing a standard 20-minute EEG containing 23 signals of frequency 250 Hz by 4600 values instead of the initial 1150000 points.

Similarity matrix computation

We only compare EEGs that have at least a subset of channels with the same labels (the channel labels corresponding to the electrodes between which a difference of electric potential i.e tension is mesured). When two EEGs don't have any channels (except the annotations channel) in common, the similarity measure between them is set to 1 (as the farther (resp. closer) the distance between two EEGs, the higher (resp. lower) and the closer to 1 (resp. closer to 0) the similarity measure).
When two EEGs are compared, their matching pairs of feature vectors (i.e vectors made of the fractal dimensions computed for each signal) do not necessarily have the same dimension. In this case, the vector of highest dimension is approximated by a histogram and the m most frequent values according to the histogram (m being the dimension of the shortest vector) are taken as representatives of that vector. Then, the distance between the two feature vectors is approximated by the distance between the shortest feature vector and the vector formed with the m most frequent values of the longest vector.

The similarity measure between two EEGs is given by:

$$\sum_{i=1}^{N} \frac{1}{N} \frac{d(ch_i^{EEG_1}, ch_i^{EEG_2} - d_{min})}{d_{max} - d_{min}} \tag{5.8}$$

where N is the number of EEG channels, $d(ch_i^{EEG_1}, ch_i^{EEG_2})$ the distance between the fractal dimensions extracted from channels with the same label in the two EEGs compared and d_{min} and d_{max} respectively the minimum and maximum distances between two EEGs in the analysed set.

We choose as metrics (d) the Euclidean distance and the normalized mutual information.

5.3.5 Description of the dataset and experiments on the fractal dimension-based similarity measure

All experiments are run on a server whose characteristics are specified in Table 5.1(a).

We apply fractal interpolation (as described in section 5.3.4) on a set of 476 EEGs[6] whose characteristics are summarized in Tables 5.1(b) and 5.1(c).

Before computing EEG distances and clustering EEGs, the EEG files on which interpolation has been applied and for which the diagnosis conclusion is either unknown or known to be abnormal without any further details are discarded. This means that the distance computation and clustering steps are performed on a subset of 328 files of the original 476 files. The similarity matrice obtained is a 328×328 matrix. The files contained in the subset chosen for clustering can be separated in 4 classes as described in Table 5.1(d). Figure 5.2 shows the plot of the envelope intensity versus the index for the euclidean-distance-based similarity measure and the plot of the envelope intensity versus the index for the mutual-information-based similarity measure. The plot for the Euclidean-distance based similarity matrix exhibits 2 distinct regions whereas the plot for the mutual-information based similarity matrix exhibits 4 distinct regions. We therefore cluster the data first in 2 different clusters using the Euclidean-based similarity matrix and then in 4 clusters using the mutual-information based matrix: this means that the mutual information-based measure yields the correct number of clusters while the Euclidean distance-based similarity measure isn't spread enough to yield the correct number of clusters.

[6] unprocessed and unnormalised

OS	Processor	RAM	Programming language
openSUSE 10.3(x86-64) (kernel version 2.6.22.5-31)	Intel® Quad-Core Xeon® E5420@2.50GHz	32GB	Python 2.6

(a) Characteristics of the server used in the experiments

Number of files	Minimum EEG duration	Maximum EEG duration	Number of files of duration <15mn	15 to 30 mn	>30mn	Minimum file size	Maximum file size	Signals frequency
476	1mn50s	5h21mn	40 (8.4%)	260 (54.6%)	176 (37%)	1133kB	138MB	250Hz

(b) Characteristics of the dataset used for fractal interpolation

Number of signals	Number of files	Percentage of files
2	1	0.2
12	1	0.2
13	2	0.4
19	13	2.7
20	63	13.2
23	386	81.1
25	7	1.5
28	3	0.6

(c) Number of signals per EEG file in the dataset used to test fractal interpolation

EEG class	Number of EEG files	Percentage of EEG files
normal	195	59.5%
epilepsy	64	19.5%
encephalopathy	31	9.5%
brain damage (i.e vascular damage, infarct or ischemia)	34	10.4%

(d) Characteristics of the dataset used for similarity computation and clustering

Table 5.1: Server and EEG test file characteristics

We compare the performance of the IFS-based similarity measure with an AR[7]-based similarity measure inspired from [101]:

- An AR model[8] is fitted to each of the signals of each of the EEG files considered (at this stage 476). The order of the AR model fitted is selected using the AIC criterion. The order is equal to 4 for our dataset.
- The LPC cepstrum coefficients are computed based on the AR model fitted to each signal using the formulas given in [101]. The number of coefficients selected is the PGCD of the number of points for all signals from all files.
- The Euclidean distance between the computed cepstral coefficients as well as the mutual information between the computed cepstral coefficients are computed in the same way as with the fractal dimension-based distances for the subset of 328 files for which the diagnosis are known. The resulting similarity matrices (328×328 matrices) are used to perform k-medoid clustering.

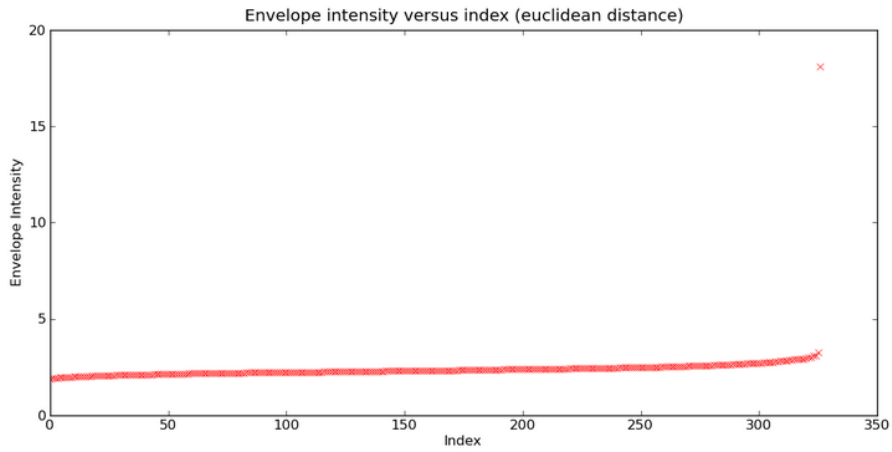
We then use the computed similarity matrices to cluster the EEGs using the method described in section 5.3.4.

5.3.6 Results of the experiments on the fractal dimension-based similarity measure

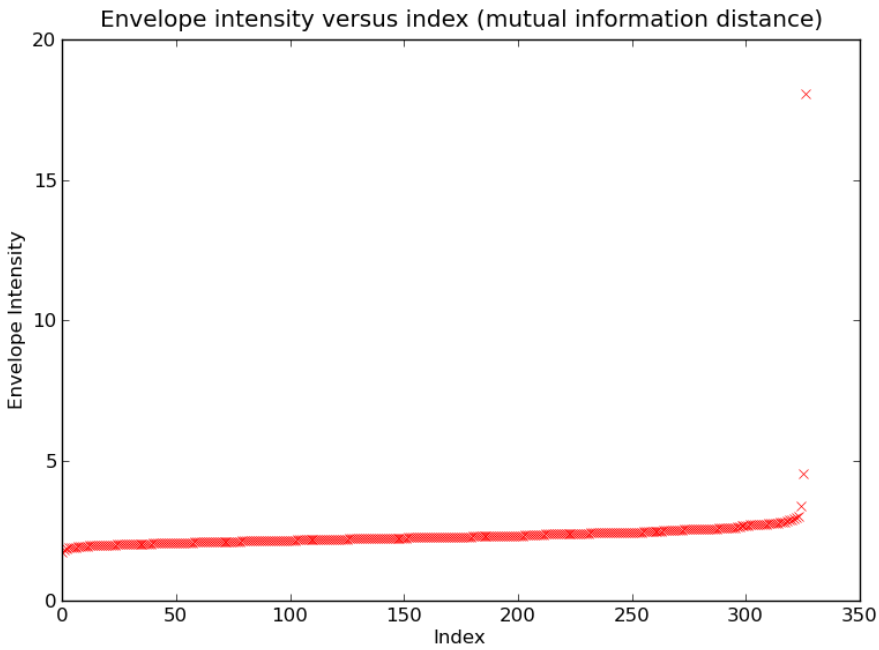
Figure 5.3 illustrates the relation between the duration of the EEG and the time it takes to interpolate EEGs. It shows that the increase of the fractal interpolation time with respect to the interpolated EEG's duration is less than linear. In comparison, AR modeling execution times increase almost linearly with the EEG duration. Therefore, fractal interpolation is a scalable method and is more scalable than AR modeling. In particular, the execution times for files of durations between 15 and 30 minutes are between 8.8 seconds and 131.7 seconds, that is execution times between 6.8 to 204.5 times lower than the duration of the original EEGs. Furthermore, the method doesn't impose any condition on the signals to be compared as it handles the cases where EEGs to be compared

[7] autoregressive model

[8] We use an AR model instead of an ARIMA model since it is argued in [101] that "for every ARIMA model there exists an equivalent AR model, that can be obtained from the ARIMA model by polynomial division" and the LPC cepstrum of the time series is, subsequently (in [101]), computed based on that equivalent AR model



(a) Euclidean distance-based matrix



(b) Mutual information-based matrix

Figure 5.2: Envelope intensity of the dissimilarity matrices

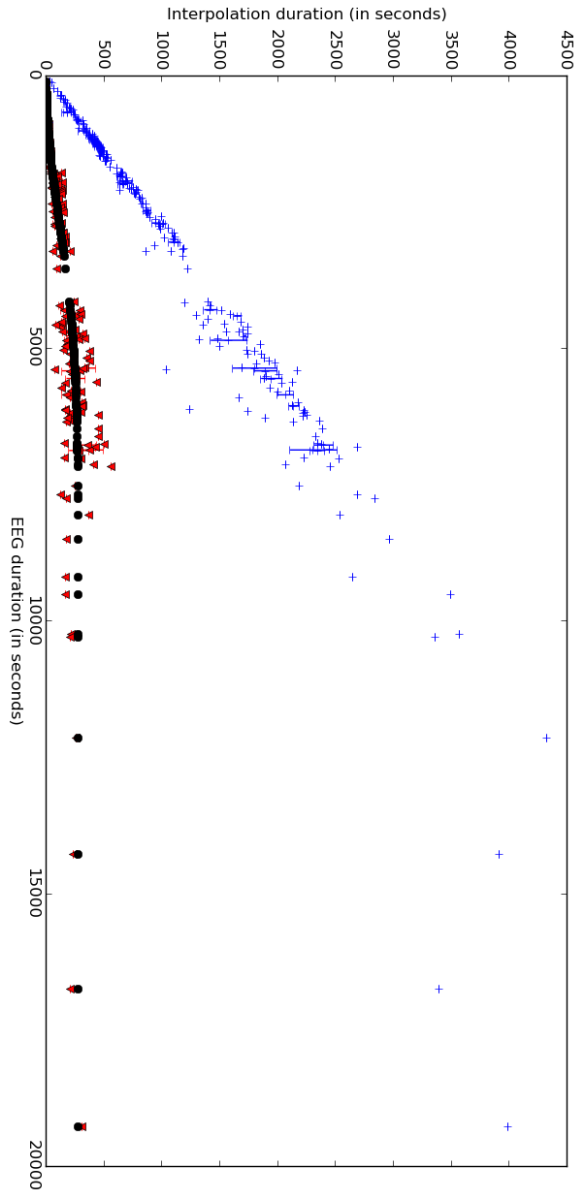


Figure 5.3: Execution times of the fractal interpolation in function of the EEG duration compared to the AR modeling of the EEGs. The red triangles represent the fractal interpolation execution times and the blue crosses the AR modeling execution times. the black stars the fitting of the fractal interpolation measured execution times with function $1.14145161064 * (1 - \exp(-(0.5 * x)^{2.0})) + 275.735500586 * (1 - \exp(-(0.000274218988011 * (x))^{2.12063087537}))$ using the Levenberg-Marquardt algorithm

Table 5.2: Specificity and sensivity of the EEG clusterings

	Specificity	**Sensitivity**
normal EEG	0.312	0.770833333333
abnormal EEG	0.770833333333	0.312

	Specificity	**Sensitivity**
normal EEG	0.297752808989	0.657534246575
epilepsy	0.65564738292	0.183006535948
encephalopathy	0.838709677419	0.051724137931
brain damage	0.818713450292	0.114285714286

have no or limited common channels and have signals of different lengths. Moreover, fractal interpolation doesn't require model selection as AR modeling does, which considerably speeds up EEG interpolation. Moreover, with our dataset, the computation of the Euclidean distance between the cepstrum coefficients calculated based on the EEGs AR models leads to a matrix of NaN (i.e not a number)[9]: the AR modeling method is therefore less stable than the fractal interpolation-based method.

Table 5.2 summarises the clustering results for all similarity matrices. The low sensitivity obtained for the abnormal EEGs (epilepsy,encephalopathy,brain damage) can be explained through the following reasons:

- most of the misclassified abnormal EEGs are EEGs representing mild forms of the pathology they point to therefore their deviation from a normal EEG is minimal
- most of the misclassified abnormal EEGs (in particular for epilepsy and brain damage) exhibit abnormalities on only a restricted number of channels (localised version of the pathologies considered). The similarity mea-

[9]The same happens when the mutual information is used instead of the Euclidean distance (all programs are written in Python 2.6)

sures, giving equal weights to all channels, are not sensitive enough to abnormalities affecting one channel.

About 76% of the normal EEGs are well classified. The remaining misclassified EEGs are misclassified because they exhibit artifacts (EMG and eye blinks in particular) and/or age-specific patterns and/or sleep-specific patterns that distort the EEGs significantly enough to make the EEGs seem abnormal. Filtering artifacts before computing the similarity measures (for example with ICA[10] and incorporating metadata knowledge (eg age associated with the EEG, type of sequence eg sleep, photostimulation) in the similarity measure would improve the clustering results.

5.3.7 Discussion

In this section, we considered the problem of defining a similarity measure for EEGs that would be generic enough to cluster EEGs without having to build an exponential number of disease-specific classifiers. We use fractal interpolation followed by fractal dimension computation to define a similarity measure. Not only does the fractal interpolation provide a very compact representation of EEGs (which may be used later on to compress EEGs) but it also yields execution times that grow less than linearly with the EEG duration and is therefore a highly scalable method. It is a method that can compare EEGs of different lengths containing at least a common subset of channels. It also overcomes several of the shortcomings of an AR modeling-based measure as it doesn't require model selection and is more stable and scalable than AR modeling-based measures. Furthermore, the mutual-information based measure is more sensitive to the correct number of clusters than the Euclidean distance-based one. It was also shown that the shortcomings of the similarity measure when it comes to clustering abnormal EEGs may be overcome through pre-processing the EEGs before interpolation to remove artifacts, tuning the weight parameters in the measure to account for small localised abnormalities and incorporating qualitative metadata knowledge to the measure.

[10] independent component analysis

5.4 EEG similarity search with fractal-based similarity measure

Section 5.3 demonstrated that relying on a feature-based similarity measure to discriminate between EEGs was a viable approach as it didn't require any stationarity assumption on the EEG, could be used to compare EEGs of different lengths provided they contain a set of channels in common and still gave promising results without any pre-processing.
It also showed that the fractal-dimension was a particularly suitable feature to discriminate between normal EEGs and abnormal ones. The fractal dimension was however not that good at discriminating between different types of abnormal EEGs. Therefore, in this section, on top of evaluating the performance of similarity search relying on a fractal-dimension based similarity measure, we will also evaluate the performance of similarity search relying on two other feature-based similarity measures, a spectral entropy-based one and a high/low frequency ratio based one (both introduced in Chapter 3). In this section, the fractal dimension is computed using the Petrosian method (see Section 5.2.5 for details).
Section 5.3 also highlighted the fact that EEGs are highly context-dependent and age-dependent and only EEGs of patients of the same age group recorded in the same context can be compared meaningfully. Consequently, in this section, we will only compare EEG segments recorded in similar contexts on adult patients. The principle of similarity search itself does not change because of this restriction. This restriction only ensures that we compare what is comparable.
We first explain what types of requests are covered by similarity search (Section 5.4.1) and what constraints apply to EEG similarity search (Section 5.4.2). We then proceed to explain the principle of the similarity search approach (Section 5.4.3), detail every step of the approach (Sections 5.4.4 and 5.4.5) and define the metrics used to evaluate the performance of the similarity search approach (Section 5.4.7). We end with describing our experimental setup and experiments (Section 5.4.6) and interpreting the experimental results (Section 5.4.8).

5.4.1 Types of user requests covered by similarity search

Similarity search in EEG data essentially covers two types of requests:

1. given a query EEG segment q (that represents a particular EEG event),

find EEG segments similar to q in a repository of EEG segments (each representing a particular EEG event)

2. given a query EEG segment q, find whole EEGs that contain q at least once.

The first type of request can be used for diagnosis by comparison. It can for example find labeled EEG segments from past patient cases that are similar to a suspicious EEG segment found in the EEG of a patient to be diagnosed, thus suggesting possible labels for the suspicious EEG segment.

The second type of request can be used to retrieve EEGs that contain the query and that point to a certain known diagnosis. For instance, if the query is a segment containing an epileptiform discharge, similarity search would retrieve EEGs that contain epileptiform discharges and are known to be pointing to a diagnosis of epilepsy because of they contain epileptiform discharges.
The second type of request may also be used when checking the impact of medication/treatment on a patient, evaluating the evolution of a particular disease or determining whether a patient presents signs of a suspected disease, even if the number of EEG segments/EEGs compared in this case would be low. For instance, if a patient is suspected to have a particular form of epilepsy, one could try and check for the presence of epileptiform discharges in the patient's recorded EEG. Another example is that of a patient with known epilepsy and whose previous EEG contained several instances of epileptiform discharge. By trying to check whether this patient's most recent EEG still contains epileptiform discharges, one can evaluate how the patient's epilepsy has evolved and whether the medication the patient is given is effective.

5.4.2 Similarity search constraints

When computing the similarity between two EEG segments, several constraints have to be taken into account.

Constraint 1: Two EEG segments showing the same clinical pattern may not have the same duration

An epileptiform discharge can last from less than 70ms to hundreds of milliseconds (depending on whether it consists in a spike, a sharp wave, a spike wave

complex or several spike wave complexes). Therefore any similarity measure between EEG segments should not be sensitive to scaling.

Constraint 2: Even if a pattern occurs in two segments to be compared, it may occur at different points of the segments to be compared.

This means in practice that the similarity measure between two EEG segments should not be sensitive to shifting.

Constraint 3: An EEG segment may contain more than one pattern, some of which may be partial

For example, an EEG segment may contain part of an artifact pattern as well as normal eyes closed patterns.

Constraint 4: The scale of electric potentials recorded in an EEG depends on the patient

To compensate for the inter-patient variability, the values of the EEG segments may need to be normalized before feature or similarity measure computation.

Constraint 5: The interpretation of EEG patterns heavily depends on the context of recording and the age of the patient

Some of the patterns recorded in children may be abnormal in adults and vice versa. And some patterns recorded in sleep may be considered pathological if seen in the EEG of an awake patient. Therefore, one must ensure when comparing two EEG segments that they have been recorded in similar contexts and on patients of the same age group. Otherwise, if checking whether the patterns of a query segment are to be found in a whole EEG, one has to extract only the EEG segments recorded in the same context as the query segment and then compare them.

5.4.3 Principle of similarity search

Similarity search is done in four steps, summarized in Figure 5.4:

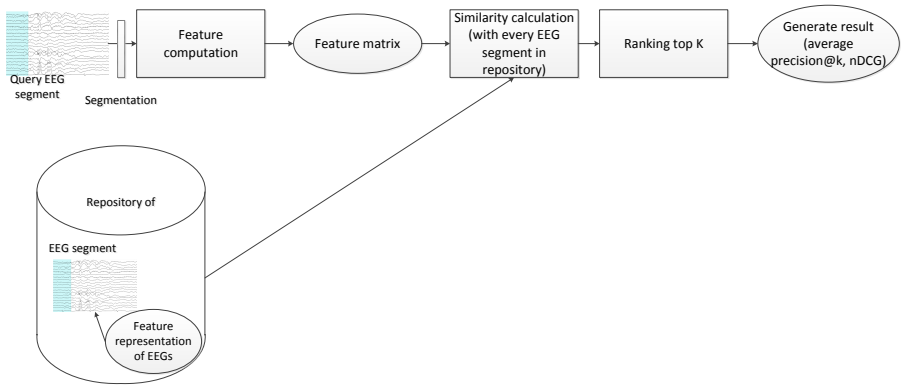


Figure 5.4: Principle of the similarity search approach

1. represent both the query EEG segment Q and the set of EEG segments to which it has to be compared/EEGs in which it must be found as feature matrices (fractal dimension matrices or entropy matrices or high/low frequency ratio matrices). See Section 5.4.4 for details.
2. with Q and the EEG segments/EEGs form the repository represented as feature matrices, compute the similarity between Q and the EEG segments/EEGs. See Section 5.4.5 for details.
3. rank the EEG segments/EEGs by their similarity to Q
4. return the top 10 EEG segments/EEGs by similarity.

5.4.4 Details on EEG segmentation and feature computation

EEG segments are transformed into feature matrices following two approaches.

First approach: Segmentation in fixed-size windows

As in Section 5.3, each channel in the EEG segment is divided in non-overlapping windows. The size of the window is chosen to be 200ms (i.e 50 points in EEG whose signals have a frequency of 250Hz).
The features are then computed for each EEG channel on each window.

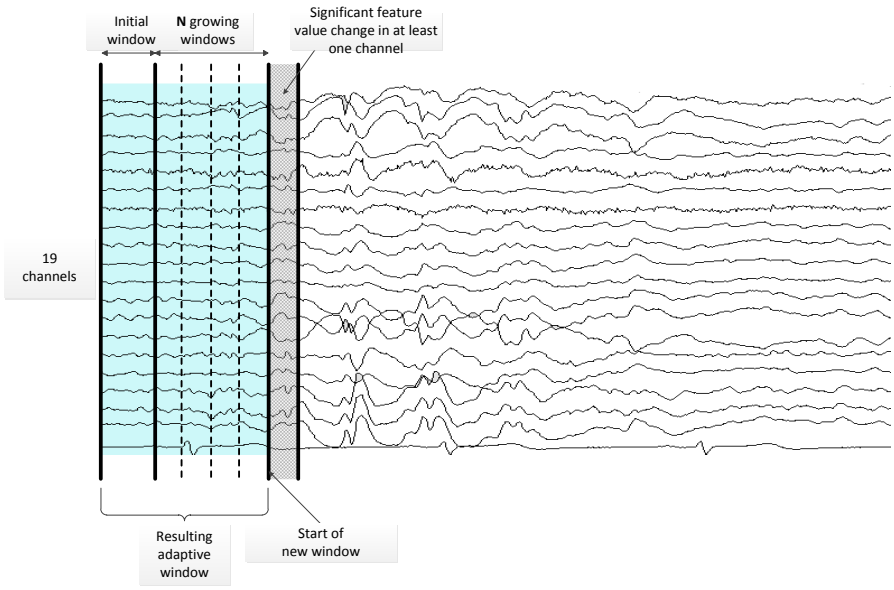


Figure 5.5: Principle of the adaptive segmentation

Second approach: Adaptive segmentation

Figure 5.5 summarizes this approach.
In this approach, we want to split each EEG segment into non-overlapping windows such that the feature values computed over subwindows of each window are (approximately) constant. In practice, this means that, for one EEG channel at least, the difference between the feature values computed on two adjacent windows falls above a certain threshold value (here, we choose that threshold value).
We first compute our feature of interest for each EEG channel on the first 100ms of the EEG segment (i.e on the first 25 points of the segment) and grow that initial 100ms window by 100ms until the difference between the feature computed on the initial window (the first 100 ms window plus N additional 100ms segments) and the feature computed on the grown window (the first 100 ms window plus $N+1$ additional 100ms segments) exceeds the threshold value of 5% for at least one EEG channel. The new window boundary is set to the start

of window $N + 1$ and the previous procedure is repeated until the entire EEG segment has been split.
The windows obtained with this approach may not have equal sizes.

Result of EEG segmentation and feature computation approaches

Both segmentation cum feature computation approaches transform the original EEG (assimilable to a matrix of dimension $N_p \times m$ with N_p the number of points per EEG channel and m the number of EEG channels) into a matrix of dimension $n \times m$ with $n << N_p$. For instance, for the fixed window approach, an EEG segment of 600ms recorded using the International 10/20 System would be transformed to an 3×19 matrix instead of the original 150×19 matrix.

5.4.5 Details on the similarity measure

Based on the constraints outlined in Section 5.4.2, the Euclidian distance and the dynamic time warping distance are not good choices of similarity metrics for our problem. The Euclidian distance supposes both sequences to be compared to be aligned so that similar patterns occur at the same time and it is also sensitive to scaling. The dynamic time warping may be insensitive to scaling but it can produce counterintuitive alignments and does not allow for gaps between patterns.
Based on this and the constraints outlined in 5.4.2, two EEG segments may be considered similar if some of their feature sequences overlap. The longer the feature overlap the more similar both EEG segments are, and the more occurrences of the overlap the more similar the segments.
Suppose we have a query EEG segment Q and EEG candidate E both represented by a feature matrix such that:

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1N_p} \\ q_{12} & q_{22} & \cdots & q_{2N_p} \\ \vdots & \vdots & \ddots & \vdots \\ q_{1m} & q_{2m} & \cdots & q_{mN_p} \end{pmatrix} \tag{5.9}$$

and

$$E = \begin{pmatrix} e_{11} & e_{12} & \dots & e_{1P} \\ e_{12} & e_{22} & \dots & e_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ e_{1m} & e_{2m} & \dots & e_{mP} \end{pmatrix} \tag{5.10}$$

Q is of dimension $m \times N_p$ and E of dimension $m \times P$. Note that P and N_p are not necessarily equal. The similarity between Q and E is computed as follows:

- for each row vector of Q, find the subsequences in the corresponding channel in E that are contained in Q and compute the mean subsequence length. Denote this value as l_i where $i \in [1, m]$. We also compute the length standard deviation std_i. Also store the number of subsequences found o_i. The result is three vectors of dimension m: $L = [l_1 l_2 ... l_m]$ and $STD = [std_1 std_2 ... std_m]$ and $O = [o_1 o_2 ... o_m]$
- compute the average of L, the average of STD and the average of O

A subsequence of E is considered to be contained in Q if the difference between each of its elements and their corresponding elements in Q falls below a threshold (set to 5% in this case). The higher the average of L and average of O and the lower the average of STD, the more similar E is to Q. When ordering segments by similarity to Q, the EEG segments are first sorted by average L, then by average STD and finally by average O. Average Ls and average Os are sorted from highest to lowest while average STDs are sorted from lowest to highest.

5.4.6 Experimental setup

The characteristics of the server used for the experiments are outlined in Table 5.3(c).
The dataset used for the experiments is described in Tables 5.3(d) and 5.3(e). We use a set of 100 segments extracted from EEGs recorded in a clinical setting at the Medisch Spectrum Twente (MST), Enschede. All EEGs from which segments were selected for inclusion in the dataset were recorded on 23 distinct adult patients in a hospital setting following the International 10/20 System with Ag/AgCl electrodes and using a common average reference. Only the 19 channels common to all EEGs are kept for calculations, with each channel

OS	Processor	RAM	Programming language
openSUSE 12.3 Milestone 2(x86-64) (kernel version 3.6.3-1-desktop)	AMD Opteron Processor 4226 2 processors (6 cores per processor)	64GB	Python 2.7.3 with joblib 0.8.3-r1

(c) Characteristics of the server used in the experiments

Datasets	Number of files	Median EEG duration	Minimum EEG duration	Maximum EEG duration	Number of files of duration <2s	2 to 10 s	>10s	Signals frequency
whole EEG segments dataset (dataset 1)	100	1.8s	600ms	201s	52 (52%)	23 (23%)	25 (25%)	250Hz
queries dataset (subset of dataset 1)	40	1.1s	600ms	201s	28 (70%)	3 (7.5%)	9 (22.5%)	250Hz

(d) Characteristics of the datasets

EEG segment label	Dataset 1		Queries dataset	
	Number of EEG segments	Percentage of EEG segments	Number of EEG segments	Percentage of EEG segments
normal eyes closed	24	24%	10	25%
eye blinks	21	21%	10	25%
sharp wave	21	21%	10	25%
spike wave complex	23	23%	10	25%
encephalopathy	11	11%	0	0%

(e) Types of EEG segments contained in the dataset

Table 5.3: Server and EEG test file characteristics

sampled at 250Hz. All segments were labeled manually (sharp wave segments and spike wave complex segments in particular were labeled by a trained EEG interpreter): the label given to a segment containing at least one of the events of interest (i.e eye blink, sharp wave, spike wave complex or normal eyes closed pattern) was the name of the event of interest. The dataset also contains segments extracted from patients with diverse types of encephalopathy: those seg-

ments are not used in the queries dataset because the exact events they represent are not known, only the final diagnosis derived from the EEGs they appear in are known. We call this dataset "dataset 1".
Fourty segments (10 for each event of interest) are randomly selected from this dataset to form the queries dataset. The events of interest are: normal eyes closed segment, sharp wave, spike wave complex and eye blinks. Note that both sharp waves and spike wave complexes are subtypes of epileptiform discharges thus very closely related and similar.

Our experiment consists in computing the similarity between each of the 40 queries in the queries dataset and each of the EEG segments from dataset 1 that are not the query segment. Therefore, each of the query segments is compared to 99 candidate EEG segments, with a total number of 3960 comparisons for all queries. For each query, a ranking of the candidate EEG segments is established based on their similarity with the query segment. The average precision@10 (see Section 5.4.7) is computed for each query and then averaged over all queries corresponding to a label. The average precision@10 is used as an overall measure of the quality of the top 10 results returned by similarity search. The normalized discounted cumulative gain (under binary and ternary assumptions) (see Section 5.4.7) is also computed for all queries averaged over all queries corresponding to a label. Under the ternary assumption, retrieved segments with the same label as the query are assigned a relevance of 2, the spike wave complex pattern is considered partially relevant to the sharp wave pattern and is assigned a relevance of 1 with regards to this pattern (and vice versa) while all the other types of segments are assigned a relevance of 0. The normalized discounted cumulative gain is used as a measure for the quality of the ranking, in other words it is used to check whether the similarity search approach manages to return the most relevant results at the top of the ranking. We derive our conclusions on the usefulness of the EEG similarity search technique based on the results of both measures combined.
The results of the experiments are shown in Tables 5.4, 5.5 and 5.6.

5.4.7 Performance metrics

Average precision@k

Given a query and the list of candidate EEG segments ranked according to their similarity to the query, we choose the top k candidate EEG segments in the list and compute the percentage of relevant candidate EEG segments (i.e

the percentage of candidate EEG segments having the same label as the query) in the reduced list. For example, if the query segment is labeled epileptiform discharge and there are 4 EEG segments with label epileptiform discharge in the top 10 candidate EEG segments, then the average precision@10 for this query is equal to 40%.
In the following experiments, we choose a value of 10 for k.

Normalized discounted cumulative gain

The normalized discounted cumulative gain assesses the quality of a ranking [110, 111, 112].
The first assumption made for this measure is that, for a given query, out of two rankings of candidate EEG segments with the same average precision@k, the best one is the one that has relevant EEG segments higher on the list than the other. For example, given a query segment labeled epileptiform discharge and two ranked lists of candidate EEG segments for which the average precision@k is equal to 0.3, the list in which the EEG segments labeled epileptiform discharge are in the first, second and third positions(rank) is considered better than the the list in which the EEG segments labeled epileptiform discharge are in the fourth, fifth and sixth position(rank).
The second assumption made for this measure is that highly relevant candidate segments are more useful than partially relevant segments, which are in turn more useful than irrelevant segments. For example, given a query segment labeled 'epileptiform discharge' and two ranked lists of candidate segments with the same average precision@k, the list that contains partially relevant segments (such as segments showing other epileptiform patterns) on top of the segments labeled epileptiform discharge is more useful than the list that contains only irrelevant segments (eg. normal eyes closed segments) on top of the segments labeled epileptiform discharge. This means that every type of candidate EEG segment has to be assigned a relevance with regards to the query.
We first compute the normalized discounted cumulative gain under a binary assumption to simplify calculations. What we mean by binary assumption is that we deem as irrelevant every segment that has a different label from the query segment label and assign a relevance of 0 to every such segment and a relevance of 1 to every segment that has the same label as the query. Doing this simplifies the computations at the risk of underestimating the quality of a ranking since some EEG patterns may be partially relevant with regards to other EEG patterns (eg. epileptiform discharge and spike wave complex patterns since both are epileptiform patterns).

The discounted cumulative gain for a query is then computed as follows:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{5.11}$$

with rel_i the relevance of the i-th ranked segment and i the candidate segment rank. The normalized discounted cumulative gain is then computed as follows:

$$nDCG_k = \frac{DCG_k}{IDCG_k} \tag{5.12}$$

with $IDCG_k$ the maximum discounted cumulative gain for the average precision@k of the ranking on which DCG_k was computed. $IDCG_k$ corresponds to the discounted cumulative gain of the ranking in which all the segments with the same label as the query are at the top of the list (eg for an average precision@10 of 0.4, all four relevant segments occupy the ranks from 1 to 4 with $displaystyle IDCG_k = \sum_{i=1}^{4} \frac{1}{\log_2(i+1)} \approx 2.562$).
We also compute the nDCG under ternary assumption i.e a relevance of 2 is assigned to highly relevant segments with regards to the query, a relevance of 1 is assigned to partially relevant segments with regards to the query and finally a relevance of 0 is assigned to irrelevant segments with regards to the query. The discounted cumulative gain remains defined as in Equation 5.11 and the normalized discounted cumulative gain defined as in Equation 5.12. The only change in calculation with this ternary assumption is the calculation of $IDCG_k$. In this case, the $IDCG_k$ corresponds to the discounted cumulative gain of the ranking in which all the segments are sorted by relevance from the highest to the lowest relevance (eg for an average precision@10 of 0.4 and with 3 partially relevant segments present, the $IDCG_k$ would be computed for the list [2 2 2 2 1 1 1 0 0 0] and would be equal to $displaystyle \sum_{i=1}^{4} \frac{3}{\log_2(i+1)} + \sum_{i=5}^{7} \frac{1}{\log_2(i+1)} \approx 8.761$).
Note that in a perfect ranking, the $IDCG_k$ and the DCG_k would be equal and the normalized discounted cumulative gain equal to 1.0. The normalized discounted cumulative gain values lies between 0 (worst possible ranking) and 1 (perfect ranking).

5.4.8 Results and discussion

Tables 5.4, 5.5 and 5.6 show the experiments' results. All average precision@10 measurements are given in Table 5.4 while the the normalized discounted cumulative gain measurements appear in Tables 5.5 and 5.6.

EEG segment class	Fractal dimension		Spectral entropy		High/low frequency ratio	
	with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment
normal eyes closed	0.97 ± 0.09	0.97 ± 0.09	0.91 ±0.14	0.93 ±0.06	0.94 ±0.12	0.85 ±0.22
eye blinks	0.15 ± 0.12	0.15 ± 0.12	0.23 ±0.11	0.25 ±0.12	0.23 ±0.11	0.24 ±0.13
sharp wave	0.16 ± 0.12	0.16 ± 0.12	0.21 ±0.09	0.19 ±0.11	0.22 ±0.098	0.22 ±0.098
spike wave complex	0.15 ± 0.12	0.15 ± 0.12	0.25 ±0.11	0.26 ±0.10	0.27 ±0.09	0.26 ±0.11

(a) Average precision@10 per feature and EEG segment class for adaptive segmentation

EEG segment class	Fractal dimension		Spectral entropy		High/low frequency ratio	
	with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment
normal eyes closed	0.97 ±0.06	0.97 ±0.06	0.97 ±0.09	0.95 ±0.09	0.90 ±0.2	0.88 ±0.2
eye blinks	0.13 ±0.1	0.13 ±0.1	0.22 ±0.1	0.24 ±0.1	0.25 ±0.12	0.24 ±0.13
epileptiform discharge	0.11 ±0.10	0.11 ±0.10	0.14 ±0.13	0.21 ±0.09	0.22 ±0.1	0.21 ±0.09
spike wave complex	0.16 ±0.13	0.16 ±0.13	0.13 ±0.13	0.22 ±0.1	0.30 ±0.08	0.29 ±0.08

(b) Average precision@10 per feature and EEG segment class for segmentation in fixed length windows

Table 5.4: Results (part 1)

Result 1: results for the normal eyes closed class

For a query segment that is an normal eyes closed segments, the top 10 most similar candidate EEG segment almost always contains 10 segments that are also normal eyes closed segments (average precision@10 ranging from 0.85 with high/low frequency ratio-based similarity measure computed on raw EEG segments split with adaptive segmentation to 0.97 for all fractal dimension-based similarity measures-see Table 5.4). The performance of the fractal dimension for normal eyes closed segments is marginally better than the other two features (normalized or not) in terms of average precision@10 (Table 5.4) while spectral entropy on raw EEG segments (nDCG of 1 for spectral entropy compared to values ranging from 0.84 to 0.94 for other features) is marginally better in terms of normalized discounted cumulative gain (with binary assumption)

EEG segment class	nDGC	Fractal dimension		Spectral entropy		High/low frequency ratio	
		with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment
normal eyes closed	Mean nDCG	0.94±0.18	0.94±0.18	0.90±0.30 (1.0±0.0)	1.0±0.0 (0.94±0.18)	0.90±0.30	0.84±0.33
	nDCG=0	0 queries	0 queries	1 query (0 query)	0 query	1 query	1 query
	0<nDCG<=0.5	1 query	1 query	0 query	0 query (1 query)	0 query	1 query
	nDCG>0.5	9 queries	9 queries	9 queries (10 queries)	10 queries (9 queries)	9 queries	8 queries
eye blinks	Mean nDCG	0.18±0.22 (0.23±0.25)	0.18±0.22 (0.24±0.25)	0.28±0.24 (0.22±0.22)	0.18±0.22 (0.22±0.22)	0.28±0.24	0.24±0.25
	nDCG=0	6 queries (5 queries)	6 queries (5 queries)	4 queries (5 queries)	6 queries (5 queries)	4 queries	5 queries
	0<nDCG<=0.5	3 queries	3 queries	4 queries	3 queries (4 queries)	4 queries	3 queries
	nDCG>0.5	1 query (2 queries)	1 query (2 queries)	2 queries (1 query)	1 query	2 queries	2 queries
sharp wave	Mean nDCG	0.12±0.24 (0.16±0.25)	0.12±0.24 (0.16±0.25)	0.12±0.24	0.12±0.24	0.12±0.24	0.12±0.24
	nDCG=0	8 queries (7 queries)	8 queries (7 queries)	8 queries	8 queries	8 queries	8 queries
	0<nDCG<=0.5	0 queries (1 query)	0 queries (1 query)	0 queries	0 queries	0 queries	0 queries
	nDCG>0.5	2 queries	2 queries	2 queries	2 queries	2 queries	2 queries
spike wave complex	Mean nDCG	0.28±0.29 (0.24±0.30)	0.28±0.29 (0.24±0.30)	0.34±0.29 (0.28±0.29)	0.38±0.26	0.38±0.26	0.34±0.29 (0.38±0.26)
	nDCG=0	5 queries (6 queries)	5 queries (6 queries)	4 queries (5 queries)	3 queries	3 queries	4 queries (3 queries)
	0<nDCG<=0.5	1 query (0 queries)	1 query (0 queries)	1 query	2 queries	2 queries	1 query (2 queries)
	nDCG>0.5	4 queries	4 queries	5 queries (4 queries)	5 queries	5 queries	5 queries

(a) Normalized discounted cumulative gain per feature and EEG segment class for both types of segmentation computed under binary assumption (in blue are the results for segmentation in fixed-size window when they differ from those of adaptive segmentation)

Table 5.5: Results (part 2)

(Table 5.5). Any of the tested features can therefore be used to retrieve seg-

EEG segment class	nDGC	Fractal dimension		Spectral entropy		High/low frequency ratio	
		with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment	with normalized EEG segment	with raw EEG segment
normal eyes closed	Mean nDCG	0.98±0.06 0.98±0.05	0.98±0.06 0.98±0.05	0.94±0.16 1.0±0.001	0.99±0.08 0.98±0.06	0.97±0.08 0.90±0.30	0.88±0.3 0.88±0.3
	nDCG=0	1 query 0 query	1 query 0 query	0 query 0 query	1 query 0 query	0 query 1 query	0 query 1 query
	0<nDCG<=0.5	4 queries 0 query	4 queries 0 query	1 query 0 query	1 query 0 query	1 query 0 query	2 queries 0 query
	nDCG>0.5	5 queries 10 queries	5 queries 10 queries)	9 queries 10 queries	8 queries 10 queries	9 queries 9 queries	8 queries 9 queries
eye blinks	Mean nDCG	0.45±0.2 0.50±0.30	0.45±0.2 0.50±0.30	0.60±0.14 0.49±0.22	0.51±0.21 0.56±0.13	0.60±0.14 0.60±0.13	0.59±0.14 0.59±0.13
	nDCG=0	1 query 2 queries	1 query 2 queries	0 queries 1 query	1 query 0 queries	0 queries 0 queries	0 queries 0 queries
	0<nDCG<=0.5	4 queries 2 queries	4 queries 2 queries	1 query 3 queries	1 query 3 queries	1 query 1 query	2 queries 2 queries
	nDCG>0.5	5 queries 6 queries	5 queries 6 queries	9 queries 6 queries	8 queries 7 queries	9 queries 9 queries	8 queries 8 queries
sharp wave	Mean nDCG	0.49±0.31 0.52±0.27	0.498±0.31 0.52±0.27	0.57±0.23 0.52±0.31	0.56±0.24 0.55±0.24	0.56±0.24 0.6±0.17	0.6±0.18 0.62±0.17
	nDCG=0	2 queries 1 query	0 queries 1 query	1 query 2 queries	1 query 1 query	1 query 0 query	1 query 0 query
	0<nDCG<=0.5	3 queries 4 queries	0 query 4 queries	2 queries 2 queries	3 queries 3 queries	3 queries 2 queries	0 query 1 query
	nDCG>0.5	5 queries 5 queries	10 queries 5 queries	7 queries 6 queries	6 queries 6 queries	6 queries 8 queries	9 queries 9 queries
spike wave complex	Mean nDCG	0.67±0.29 0.68±0.33	0.67±0.29 0.68±0.33	0.73±0.16 0.62±0.35	0.75±0.14 0.76±0.17	0.75±0.15 0.72±0.16	0.71±0.16 0.72±0.16
	nDCG=0	1 query 1 query	1 query 1 query	0 queries 2 queries	0 queries 0 queries	0 queries 0 queries	0 queries 0 queries
	0<nDCG<=0.5	1 query 1 query	1 query 1 query	1 query 1 query	0 queries 1 query	1 query 1 query	0 queries 1 query
	nDCG>0.5	8 queries 7 queries	8 queries 7 queries	9 queries 7 queries	10 queries 9 queries	9 queries 9 queries	10 queries 9 queries

(a) Normalized discounted cumulative gain per feature and EEG segment class for both types of segmentation computed under ternary assumption (in blue are the results for segmentation in fixed-size window and in black those for adaptive segmentation)

Table 5.6: Results (part 3)

ments that are normal eyes closed segments.

Result 2: results for sharp wave, spike wave complex and eye blink patterns

For all the query segments that are not normal eyes closed segments, the top 10 most similar candidate EEG segment contains from two to three segments that are of the class of the query segment (average precision@10 ranging from 0.11 to 0.30 depending on the feature, segmentation and pre-processing-Table 5.4). Both the entropy and the high/low frequency ratio (in particular the high/low frequency ratio computed after normalizing the raw EEG segments' values) perform slightly better than the fractal dimension in this case.

Regarding the sharp wave class, Table 5.5(a) shows that the similarity search fails to retrieve segments of the same class at high ranks for all but two queries. When inspecting the query segments, it turns out the sharp pattern only appears in part of the channels for all but two queries. For the two queries for which similarity search yields good results, the sharp wave pattern appears in all channels. As a consequence, because the similarity measure computes the mean of all similarities per channel without assigning different weights to different channels, the sharp wave pattern may be confused with the eye blink pattern, which is also localized on some channels though the sharp wave pattern and the eye blink pattern usually appear in different channels.
The low normalized discounted cumulative gains (under binary assumption) for the sharp wave class (see Table 5.5) may be explained by the fact that both the sharp wave pattern and the spike wave complex pattern are very closely related since they are both subtypes of epileptiform discharges. In fact, the spike wave complex pattern is composed of a spike followed by a slow wave, with the spike being similar in shape to a sharp wave but shorter in duration (typically a spike lasts between 20 to 70ms while the sharp wave lasts from 70 to 200ms). With our similarity measure quantifying the overlap between two segments, the sharp wave and spike wave complex patterns appear somewhat similar. However, retrieving spike wave complexes instead of sharp waves would still be informative since finding spike wave complexes in an EEG segment still points to a diagnosis of epilepsy (as is the case for sharp waves). When computing the normalized discounted cumulative gains (under binary assumption) (see Table 5.5), we considered spike wave patterns to be irrelevant with regards to query containing a sharp wave pattern (and vice versa). Consequently, the quality of the rankings for those two patterns may be underestimated. To verify that, we computed the normalized discounted cumu-

lative gains under ternary assumption, shown in Table 5.6(a). The normalized discounted cumulative gain values for both sharp wave and spike wave complex patterns computed under ternary assumption (Table 5.6(a)) are markedly higher than the corresponding normalized discounted cumulative gain values computed under binary assumption. This means that similarity search retrieves many partially relevant segments with regards to the 'sharp wave' class on top of retrieving some segments with sharp wave patterns in the top 10.

Result 3: Effect of the normalization of the EEG data

The results for the fractal dimension do not vary with normalization while both entropy and high/low frequency ratio perform better when the raw EEG data is normalized prior to feature and similarity measure computation.

Result 4: Effect of the segmentation

The results when using adaptive segmentation are on the whole (with the exception of a few features) are marginally better then those obtained when segmenting EEG segments into fixed size windows. This may be because the size of the adaptive window after one iteration is equal to the size of the fixed window, leading to similar segmentations in this case. Adding smaller windows per iteration in the adaptive segmentation may yield a better performance. The influence of the window size per iteration parameter on the overall similarity search performance needs to be studied further.

Directions for similarity search performance improvement

The results shown in Section 5.4.8 hint at the fact that it may be more difficult to distinguish patterns that only occur on certain channels (eg eye blinks patterns that typically occur in channels measuring the potentials at electrodes Fp1 and Fp2) since we average the similarities per channel and give the same weight to all channels in the average when computing similarity thus possibly losing localization information. An avenue for future work would be studying the weighting of the EEG channels and the incorporation of channel localization information in the similarity measure to achieve better discriminative performance for classes other than eyes closed segments.

Furthermore, no artifacts were removed form the EEG segments prior to the computations which may have affected the results: it could be useful in future to investigate the impact of applying artifact removal algorithms (for example artifact removal with ICA i.e independent component analysis) on the retrieval performance of classes other than normal eyes closed segments.
In practice, the queries submitted for similarity search are mostly segments whose patterns deviate from the normal eyes closed pattern. In this case, normalizing the query and candidate segments prior to feature and similarity measure computation as well as using features other than the fractal dimension, in particular the high/low frequency ratio, seems to be the method to follow. In future research, it may be worthwhile to study the performance of different combinations of features.
To be able to assess the retrieval performance of our similarity search approach, all experiments were done on manually labeled segments. In practice, not all segments in stored EEGs are labeled. This is a case where user feedback may be gainfully used: users would be asked to assess the relevance, with regards to their query, of the top ranked unlabeled segments in particular so that the retrieval performance for said query can be measured. User feedback could also be combined to the similarity measure as an additional term (in particular as a penalizing term for segments not similar to the query) so as to improve its performance (method shown in Chapter 4). A user study needs to be done to validate this point.
All in all, *the results for the proposed feature-based similarity search can be considered useful*: relevant and partially relevant information is present in the top 10 segments retrieved by similarity search thus easily accessible to the user. Improving the performance of the similarity search would not require a change of paradigm but only exploring variable parameters with the proposed similarity search paradigm (eg exploring new features and their combination, adding some pre-processing such as artifact removal or adding metadata information such as the channel localization of similar sequences).

5.5 Conclusion

In this chapter, we have studied the suitability of feature-based similarity measures for EEG similarity search. We have studied three main features: the fractal dimension, the spectral entropy and the high/low frequency ratio. We have also assessed the impact of normalizing the EEG data prior to feature and similarity measure computation on the retrieval of relevant candidate EEG seg-

ments with regards to the initial query EEG segment.
The first part of the chapter (Section 5.3) proved that relying on a feature-based similarity measure to discriminate between EEGs was a valid approach: not only could it be applied without any stationarity assumption on the EEG, but it could also be used to compare EEGs of different lengths (provided they contain a set of channels in common) and still gave promising results without any pre-processing. It showed however that other features besides the fractal dimension may be needed to discriminate between different types of abnormal EEG patterns.
Section 5.4 extended the work done in Section 5.3 and proposed an approach to EEG similarity search based on several feature-based similarity measures. The performance of the similarity search for 4 types of EEG patterns, some of which are closely related, was studied and the influence of several parameters on similarity search performance was considered, among others the influence of EEG segmentation and normalization.
It was shown that the fractal dimension is not affected by normalization and separates normal eyes closed segments very well from other patterns but is less effective at discriminating between patterns that are not normal eyes closed segments. Spectral entropy and high/low frequency ratio are better at distinguishing between patterns that are not normal eyes closed segments and their performance is enhanced by applying normalization prior to feature and similarity measure computation. Though perfectible, the performance achieved with the current similarity search approach proves this approach to be **already useful** since the current approach retrieves relevant and partially relevant information with regards to the EEG segment query in the top 10 results list.
This study is a first step towards designing a similarity measure suitable for EEG similarity search. The proposed feature-based similarity measure can compare EEG segments of different lengths provided they contain at least a common subset of channels. The proposed similarity measure is also insensitive to shifting and based on a compact representation (through feature matrices) of the EEG segments to compare. If the feature representations of EEG segments are stored alongside the original EEG segment values, then this may make the computation of the similarity between EEG segments scalable.
In future work, it would be useful to investigate other features and in particular combinations of features and assess their discriminative performance with regards to patterns that are not normal eyes closed segments. It would also be worthwhile to evaluate the impact of artifact removal prior to feature and similarity measure computation on the retrieval of classes other than normal eyes closed segments. Another topic of research would be incorporating channel

localization in the similarity measure so as to differentiate between localized patterns.
A last subject of research may be assessing the relevance of retrieved segments through user feedback (for example with the model introduced in Chapter 4), in particular the relevance of unlabeled segments and adding this relevance feedback with the proposed similarity measure as an additional boosting/penalizing term to improve the retrieval performance.

Acknowledgments

We would like to thank Michiel Punter (scientific programmer at the Technical Medicine department) for all his precious help in programming and comments on the algorithms presented in Section 5.3. We would also like to thank Mena Badieh Habib Morgan, postdoctoral researcher at the Database Group, for helping draw some of the figures included in this chapter.
In addition, we would like to thank Dr. Ouafa Mouti, neurologist in Rabat(Morocco), for her crash course on EEG basics and for providing all the resources to help understand EEGs better. We would like to thank Marleen Cloostermans, Esther ter Braack and Shaun Lodder, PhD students at the Technical Medicine Department for their explanations on EEGs and signal processing.

CHAPTER 6

Conclusions

6.1 Summary of the problem: the misdiagnosis problem

"Errare humanum est" [1]. Clinicians, however smart, caring and meticulous they may be, are only human, all too human. So they are bound to occasionally make mistakes, diagnosis errors (i.e delayed/missed or wrong diagnosis) in particular. With a prevalence of misdiagnosis of 15% in most areas of medicine ([2]), of which about 32% are due to errors in clinician assessments, misdiagnosis is a major if overlooked problem which seems to have systemic roots rather than just being the problem a few isolated "bad apples". Assigning blame to individual health practitioners for diagnosis failures won't solve the systemic issues and will only ensure that the same avoidable mistakes are repeated. Understanding the root causes of systemic failures leading to misdiagnosis, however, can help devise solutions that minimize the occurrence of misdiagnoses and/or their impact since, as stated in the landmark 2000 Institute of Medicine report on medical errors,"*Errors can be prevented by designing systems that make it hard for people to do the wrong thing and easy for people to do the right thing*" ([56]). The main reason so many misdiagnoses occur is because diagnosis is a decision-making process made under constraints that may conflict with the accuracy requirement. Such constraints include the following:

- cost constraints
- clinician time/availability and energy constraints
- high data uncertainty

[1] To err is human

- and the unavoidable fragmentation of the diagnosis process (ie several agents-eg clinicians, nurses and lab technicians- involved in the diagnosis process) which is due to the necessarily finite amount of memory and knowledge healthcare practitioners possess as a result of their human condition and the amount of medical knowledge collected over the centuries.

Those constraints lead to two main misdiagnosis risks factors:

- a fragmentation of patient data, in particular history, which means that clinicians may not have all the clues and evidence needed to make the correct diagnosis, and

- the reliance on reasoning shortcuts and heuristics, which, if applied correctly, minimizes the time and effort required to reach a diagnosis but can on the contrary cause harm if they lead to making/confirming the wrong diagnosis hypothesis, interpreting diagnostic clues incorrectly or not considering other likely differential diagnoses too soon in the process.

6.2 Goals of the research

So how can we use database/data mining knowledge to support clinicians in diagnosis process and obviate the risks posed by data fragmentation and cognitive shortcuts and biases? In response to this question, we designed a medical data sharing platform with the following objectives:

1. share patient data and make it easily accessible

2. help researchers help clinicians by providing a standard trove of data, to ensure (semi)-automated medical data interpretation methods are more easily comparable and reproducible, and by providing a data processing platform

3. make it easy to browse the data with similarity search requests (useful for differential diagnosis or diagnosis by comparison)

4. combine evidence at hand to provide a set of diagnosis hypotheses and their attached likelihood

Sharing patient data to make it accessible from one place (goal 1) would help solve the problem of data fragmentation and help make sure clinicians can access all the needed relevant data when forming their diagnosis hypotheses.
And because we would allow access to the data to researchers (under conditions respecting patient privacy) so that they have standard data as well as a processing platform to run experiments on (goal 2), we would make it easier for them to develop and compare the performance of (semi)-automated medical data interpretation methods designed to alleviate clinicians' workload and let them focus on challenging cases.
Making it simple to browse the data with similarity search (goal 3) would allow researchers to easily build experimental datasets as well as allow clinicians to easily make differential diagnoses or diagnoses by comparison.
By combining all the evidence at hand (goal 4) to provide a set of diagnosis hypotheses and their attached likelihoods, several possible hypotheses would be presented to the clinician at each step. Some of the evidence used in the combination can, for example, be obtained from software that matches clinical findings against a database of medical conditions, much larger than the database held in a single clinician's memory, to determine the likelihood of a condition such as the software presented in [9]. Combining all the available evidence would help reduce the incidence and impact of cognitive biases such as premature closure, confirmation bias, zebra retreat, framing bias and diagnosis momentum as well as availability bias and representativeness bias since the clinician would be prompted to consider several diagnosis possibilities instead of just one, some of which may be zebras or diseases that have the clinical evidence at hand as atypical presentation.

6.3 Contributions to minimizing the misdiagnosis problem

To try and reach the goals outlined in the previous section (Section 6.2), four main contributions have been made in this thesis:

1. Contribution 1: a feasibility study for Hadoop as medical sharing and processing platform

2. Contribution 2: a similarity measure based on features extracted from EEG data

3. Contribution 3: a Dempster-Shafer based evidence combination framework to deal with uncertainty in incremental decision-making processes (eg diagnosis process)

4. Contribution 4: using contribution 2 to process similarity search requests on EEG data and explaining how contribution 3 may be combined to contribution 2 to improve EEG similarity search results

The feasibility study (Chapter 3) shows that technology to share data at little expense and effort already exists. And because Hadoop can handle diverse data formats, there is no need to standardize the existing data formats as long as methods to read and/or visualize them exist and are made available. In that sense, the only step needed to start sharing medical data is to start deploying Hadoop in medical institutions and transferring data to the Hadoop platform. The feasibility study in Chapter 3 also demonstrates that this platform is suitable for developing medical data interpretation methods. In the study, we showed with a representative data set (ie EEG data) that the Hadoop platform can perform one of the most computationally expensive data mining tasks (ie exhaustive search feature selection) on national scale amounts of representative data, thus proving the readiness in performance and scalability for medical data interpretation methods.
So contribution 1 allows us to mostly reach goals 1 and 2.
Contributions 2 and 4 (Chapter 5) are a first step towards serving similarity search requests targeting multidimensional time series data and are a first step towards reaching goal 3.
Contribution 3 (Chapter 4) provides the theoretical framework needed to reach goal 4.

6.4 Future work

We said earlier that all that was needed to start sharing medical data was deploying Hadoop in medical institutions. An important concern that arises before deploying the data with Hadoop is its security and privacy given the sensitive nature of the data. Research is ongoing to ensure data security on Hadoop (eg [113, 114]) but more needs to be done on this topic.
Contribution 3 was made under the assumption that all sources of evidence that provide a diagnosis hypothesis likelihood can be combined. The said diagnosis likelihood can be precisely quantified or just specified loosely and

qualitatively provided that a scheme is defined to convert the qualitative specification to a quantitative measure. As such, the evidence that can be used is varied: previous clinician conclusions', output of semi-automated medical data methods, methods that match sets of clinical findings (physical findings or test results) to corpora of medical conditions, etc. Going further, a prototype that combines the previously cited sources of evidence should be built and then experimentally validated through a user study that would in particular study how the source reliability coefficients should be set and what impacts those coefficients have.

Contribution 3 has also focused on combining likelihoods of discrete variables. Contribution 3 could be expanded by building and evidence combination/feedback model for (continuous) variables defined by probabilistic density functions, eg a patient's temperature or blood test results. Such a model may be useful in assessing the reliability of test results for example.

This study should be extended to other types of medical data in particular image data (eg MRI, CT scans). In particular, similarity search strategies should be devised for such data. There is also a need to study the integration of results from similarity search applied on different types of data.

Contributions 2 and 4 are a first step towards serving similarity search requests for EEG-like data ie multidimensional time series. To improve the EEG similarity search, the methods proposed in contributions 2 and 4 should be integrated with other mature semi-automated EEG interpretation methods such as the ones proposed in [13, 60, 54].

APPENDIX A

Proof of concept implementation

As a proof of concept, we have developed a small database storing EEGs, pieces of evidence, the mass functions associated with each piece of evidence as well as lineage and versioning. In this appendix, we provide some details about this system. Note that the system is not completely consistent with the described model in the paper, but rather with a predecessor model. Nonetheless we believe that it is illustrative of what a database implementation of the model could look like.

In Section 4.8 we argued that such a system could be implemented with a probabilistic database. Since probabilistic database prototype systems do not support all research results mentioned in literature, we decided to build our prototype not on top of an existing probabilistic database prototype, but rather in a conventional relational database (PostgreSQL 8.4). The system is heavily inspired by the probabilistic database prototype Trio [81, 79, 80].

The schema of our proof of concept system contains the following tables.

- `eeg_certain`
- `eeg_uncertain`
- `eeg_certain_versioning`
- `eeg_uncertain_versioning`
- `eeg_lineage`

The `eeg_certain` table contains the certain attributes of EEGs such as filename, fileidentifier and xid (x-tuple id). The `eeg_uncertain` table, meanwhile, contains attributes with uncertainty such as the diagnosis evidence associated with the EEG (`label` attribute) and the confidence value associated with that diagnosis (`conf` attribute). If an EEG has, for example, two possible diagnoses associated to it (epilepsy with a confidence p_1 and artifact with confidence $1-p_1$), then a tuple is associated with the said EEG in the `eeg_certain` table and two tuples are associated with the said EEG in the `eeg_uncertain`

Table A.1: Example of EEG metadata storage

Table `eeg_certain`

xid	filename
1	EEG1.edf

Table `eeg_uncertain`

aid	xid	label	conf
1	1	epilepsy	p_1
2	1	artifact	$1 - p_1$

table (one for each alternative) as shown in Table A.1. Each alternative is identified with the unique identifier `aid` which is linked to the `eeg_certain` table with the `xid` identifier.

Versioning support is given with the versioning tables `eeg_certain_versioning` and `eeg_uncertain_versioning`. These tables keep track of modifications to the source tables `eeg_certain` and `eeg_uncertain`. In the proof of concept system, we identify each different version with a version identifier. In this way, the versioning tables contain all but the current version of the source information while the source tables contain only the current version.

The `eeg_lineage` table stores the provenance of each of the tuples stored in the `eeg_uncertain` table and, in particular, the provenance of the confidence values associated with each EEG diagnosis alternative recorded in that table. It records, for instance, whether a particular diagnosis alternative and its confidence value are derived, for example, by similarity of an EEG with another which has a diagnosis attached. It also records whether or not a particular diagnosis alternative has been subjected to meta-evidence.

A.0.1 Details of the tables' implementation

The following tables include various details about each of the tables contained in our proof of concept system.

Table "public.eeg_certain"

Column	*Type*
xid	integer
filename	text

Table "public.eeg_uncertain"

Column	*Type*
aid	integer
xid	integer
label	text
conf	double precision

Table "public.eeg_certain_versioning"

Column	*Type*
xid	integer
filename	text
version	integer

Table "public.eeg_uncertain_versioning"

Column	*Type*
aid	integer
xid	integer
label	text
conf	double precision
version	integer

Table "public.eeg_lineage"

Column	*Type*
aid	integer
src_aid	integer
src_tables	text
parent_xids	text
parent_aids	text
type_feedback	text
user_bpa	double precision
user_id	text
feedback_weight	double precision

Bibliography

[1] J. Dwyer, "An infection, unnoticed, turns unstoppable," *The New York Times*, July 11 2012. [Online]. Available: http://www.nytimes.com/2012/07/12/nyregion/in-rory-stauntons-fight-for-his-life-signs-that-went-unheeded.html?pagewanted=all&_r=2&

[2] M. L. G. Eta S. Berner, "Overconfidence as a cause of diagnostic error in medicine," *Am. J. Med*, vol. 121, no. 5, pp. S2–S23 (Supplement), May 2008.

[3] G. D. Schiff, O. Hasan, S. Kim, R. Abrams, K. Cosby, B. L. Lambert, A. S. Elstein, S. Hasler, M. L. Kabongo, N. Krosnjar, R. Odwazny, M. F. Wisniewski, and R. A. McNutt, "Diagnostic error in medicine: Analysis of 583 physician-reported errors," *Archives of internal medicine*, vol. 169, no. 20, pp. 1881–1887, 2009.

[4] L. L. Leape, D. M. Berwick, and D. W. Bates, "Counting deaths due to medical errors [reply]," *JAMA*, vol. 288, no. 19, p. 2405, 2002. [Online]. Available: http://dx.doi.org/10.1001/jama.288.19.2405-JLT1120-2-3

[5] O. Kostopoulou, B. C. Delaney, and C. W. Munro, "Diagnostic difficulty and error in primary care–a systematic review," *Family Practice*, vol. 25, no. 6, pp. 400–413, 2008. [Online]. Available: http://fampra.oxfordjournals.org/content/25/6/400.abstract

[6] H. Singh, T. D. Giardina, A. N. D. Meyer, S. N. Forjuoh, M. D. Reis, and E. J. Thomas, "Types and origins of diagnostic errors in primary care settings," *JAMA Internal Medicine*, vol. 173, no. 6, pp. 418–425, 2013. [Online]. Available: +http://dx.doi.org/10.1001/jamainternmed.2013.2777

[7] I. M. Benseñor, "Do you believe in the power of clinical examination? The answer must be yes!" *Sao Paulo Med J*, vol. 121, no. 6, Nov. 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/14989136

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," Mc Kinsey Global Institute, Tech. Rep., June 2011. [Online]. Available: http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx

[9] M. Segal, "How doctors think, and how software can help avoid cognitive errors in diagnosis," *Acta Pædiatrica*, vol. 96, no. 12, pp. 1720–1722, 2007. [Online]. Available: http://dx.doi.org/10.1111/j.1651-2227.2007.00480.x

[10] OECD, *Health at a Glance 2011: OECD Indicators* . OECD Publishing, 2011, ch. Medical technologies.

[11] S. G. Boodman, "Medical mystery: A persistent and pernicious headache," *The Washington Post*, January 28 2013. [Online]. Available: http://www.washingtonpost.com/national/health-science/a-persistent-and-pernicious-headache/2013/01/28/13ed703e-4d2e-11e2-950a-7863a013264b_story.html

[12] L. Hao, R. Ghodadra, and N. V. Thakor, "Quantification of Brain Injury by EEG Cepstral Distance during Transient Global Ischemia," in *Proceedings - 19th International Conference - IEEE/EMBS*, Chicago, IL., USA, Oct. 30 - Nov. 2 1997.

[13] R. B. Pachori, "Discrimination between ictal and seizure-free EEG signals using empirical mode decomposition," *Research Letters in Signal Processing*, vol. 2008, pp. 1–5, 2008.

[14] T. Bermudez, D. Lowe, and A.-M. Arlaud-Lamborelle, "EEG/ECG information fusion for epileptic event detection," in *Proceedings of the 16th international conference on Digital Signal Processing*, ser. DSP'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 824–831. [Online]. Available: http://dl.acm.org/citation.cfm?id=1700307.1700445

[15] C. Goh, B. Hamadicharef, G. T. Henderson, and E. C. Ifeachor, "Comparison of Fractal Dimension Algorithms for the Computation of EEG Biomarkers for Dementia," in *Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2005)*, Costa da Caparica, Lisbon, Portugal, June 29- July 1 2005.

[16] American Clinical Neurophysiology Society, "Guideline 8: Guidelines for recording clinical eeg on digital media," *Journal of clinical neurophysiology*, vol. 23, pp. 122–124, April 2006.

[17] American Electroencephalographic Society, "Guideline 6: A proposal for standard montages to be used in clinical eeg," *Journal of clinical neurophysiology*, vol. 23, pp. 111–117, April 2006.

[18] D. Schomer and F. da Silva, *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Wolters Kluwer Health, 2012.

[19] M. J. Aminoff, *Electrodiagnosis in Clinical Neurology*, 5th ed. Elsevier Churchill-Livingstone, 2005.

[20] M. H. Libenson, *Practical approach to electroencephalography*. Elsevier Health Sciences, 2012.

[21] T. Yamada and E. Meng, *Practical Guide for Clinical Neurophysiologic Testing: EEG*, 1st ed., ser. M - Medicine Series. Lippincott Williams & Wilkins, 2009.

[22] L. Hirsch and R. Brenner, *Atlas of EEG in Critical Care*. John Wiley & Sons, Ltd, 2010.

[23] E. Niedermeyer and F. Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Williams & Wilkins, 1999.

[24] S. Sanei and J. Chambers, *EEG signal processing*. New York: Wiley-Interscience, 2007.

[25] L. J. Greenfield, J. D. Geyer, and P. R. Carney, *Reading EEGs: A Practical Approach*, 1st ed. Lippincott Williams & Wilkins, 2009.

[26] B. Kemp, A. Värri, A. C. Rosa, K. D. Nielsen, and J. Gade, "A simple format for exchange of digitized polygraphic recordings," *Electroencephalography and clinical neurophysiology*, vol. 82, pp. 391–3, May 1992.

[27] B. Kemp and J. Olivan, "European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data," *Clinical neurophysiology*, vol. 114, pp. 1755–61, September 2003.

[28] P. N. Modur and B. Rigdon, "Diagnostic yield of sequential routine {EEG} and extended outpatient video-eeg monitoring," *Clinical Neurophysiology*, vol. 119, no. 1, pp. 190 – 196, 2008.

[29] T. E. Losey and L. Uber-Zak, "Time to first interictal epileptiform discharge in extended recording eegs," *Journal of Clinical Neurophysiology*, vol. 25, no. 6, pp. 357–360, 2008.

[30] W. Webber, B. Litt, R. Lesser, R. Fisher, and I. Bankman, "Automatic EEG spike detection: what should the computer imitate?" *Electroencephalography and Clinical Neurophysiology*, vol. 87, no. 6, pp. 364 – 373, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/001346949390149P

[31] W. Webber, B. Litt, K. Wilson, and R. Lesser, "Practical detection of epileptiform discharges (eds) in the EEG using an artificial neural network: a comparison of raw and parameterized EEG data," *Electroencephalography and Clinical Neurophysiology*, vol. 91, no. 3, pp. 194 – 204, 1994.

[32] R. D. Jones, A. Dingle, G. Carroll, R. Green, M. Black, I. Donaldson, P. Parkin, P. Bones, and K. Burgess, "A system for detecting epileptiform discharges in the eeg: real-time operation and clinical trial," in *Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE*, vol. 3. IEEE, 1996, pp. 948–949.

[33] S. S. Lodder, J. Askamp, and M. J. van Putten, "Inter-ictal spike detection using a database of smart templates," *Clinical Neurophysiology*, vol. 124, no. 12, pp. 2328 – 2335, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1388245713006895

[34] S. S. Lodder and M. J. A. M. van Putten, "A self-adapting system for the automated detection of inter-ictal epileptiform discharges," *PLoS ONE*, vol. 9, no. 1, p. e85180, 01 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0085180

[35] M. J. van Putten, "Nearest neighbor phase synchronization as a measure to detect seizure activity from scalp EEG recordings," *Journal of Clinical Neurophysiology*, vol. 20, no. 5, pp. 320–325, 2003.

[36] S. S. Lodder, J. Askamp, and M. J. A. M. van Putten, "Computer-assisted interpretation of the eeg background pattern: A clinical evaluation," *PLoS ONE*, vol. 9, no. 1, p. e85966, 01 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0085966

[37] K. Inoue, K. Kumamaru, S. Sagara, and S. Matsuoka, "Pattern recognition approach to human sleep eeg analysis and determination of sleep stages." *Memoirs of the Kyushu University, Faculty of Engineering*, vol. 42, no. 3, pp. 177–195, 1982.

[38] P. S. Jensen, H. B. Sorensen, H. L. Leonthin, and P. Jennum, "Automatic sleep scoring in normals and in individuals with neurodegenerative disorders according to new international sleep scoring criteria," *Journal of Clinical Neurophysiology*, vol. 27, no. 4, pp. 296–302, 2010.

[39] M. C. Cloostermans, C. C. de Vos, and M. J. van Putten, "A novel approach for computer assisted EEG monitoring in the adult ICU," *Clinical Neurophysiology*, vol. 122, no. 10, pp. 2100 – 2109, 2011.

[40] S. P. Finnigan, M. Walsh, S. E. Rose, and J. B. Chalk, "Quantitative {EEG} indices of sub-acute ischaemic stroke correlate with clinical outcomes," *Clinical Neurophysiology*, vol. 118, no. 11, pp. 2525 – 2532, 2007.

[41] M. J. van Putten, "The Colorful Brain: Visualization of EEG Background Patterns," *Journal of Clinical Neurophysiology*, vol. 25, no. 2, pp. 63–68, 2008. [Online]. Available: http://doc.utwente.nl/78039/

[42] J. J. Halford, "Computerized epileptiform transient detection in the scalp electroencephalogram: Obstacles to progress and the example of computerized ECG interpretation," *Clinical Neurophysiology*, 2009.

[43] G. Berrada, M. van Keulen, and M. B. Habib, "Hadoop for EEG storage and processing: A feasibility study," in *Brain Informatics and Health*. Springer, 2014, pp. 218–230.

[44] K. Wiley, A. Connolly, J. P. Gardner, S. Krughof, M. Balazinska, B. Howe, Y. Kwon, and Y. Bu, "Astronomy in the cloud: Using mapreduce for image coaddition," *CoRR*, vol. abs/1010.1015, 2010.

[45] F. Bach, H. K. Çakmak, H. Maass, and U. Kuehnapfel, "Power Grid Time Series Data Analysis with Pig on a Hadoop Cluster compared to Multi Core Systems," in *Proceedings of the 21th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2013*, 2013.

[46] H. Dutta, A. Kamil, M. Pooleery, S. Sethumadhavan, and J. Demme, "Distributed storage of large-scale multidimensional electroencephalogram data using hadoop and hbase," in *Grid and Cloud Database Management*, S. Fiore and G. Aloisio, Eds. Springer, 2011, pp. 331–347.

[47] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *OSDI*. USENIX Association, 2004, pp. 137–150.

[48] R. J. Chansler, "Data availability and durability with the Hadoop Distributed File System," vol. 37, no. 1, pp. 16–22, 2012. [Online]. Available: https://www.usenix.org/publications/login/february-2012/data-availability-and-durability-hadoop-distributed-file-system

[49] D. Borthakur, J. Gray, J. S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, R. Schmidt, and A. Aiyer, "Apache hadoop goes realtime at facebook," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, ser. SIGMOD '11. New York, NY, USA: ACM, 2011, pp. 1071–1080. [Online]. Available: http://doi.acm.org/10.1145/1989323.1989438

[50] K. E. Misulis, *Atlas of EEG, Seizure Semiology, and Management*, second edition ed. Oxford University Press, 2013. [Online]. Available: http://books.google.nl/books?id=RWQGAQAAQBAJ

[51] T. Cecchin, R. Ranta, L. Koessler, O. Caspary, H. Vespignani, and L. Maillard, "Seizure lateralization in scalp EEG using Hjorth parameters," *Clinical Neurophysiology*, vol. 121, no. 3, pp. 290–300, Mar. 2010. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00441130

[52] G. Berrada and A. de Keijzer, "An IFS-based similarity measure to index electroencephalograms," in *PAKDD (2)*, 2011, pp. 457–468.

[53] M. J. van Putten, "Extended BSI for continuous EEG monitoring in carotid endarterectomy," *Clinical Neurophysiology*, vol. 117, no. 12, pp. 2661–2666, 2006. [Online]. Available: http://doc.utwente.nl/63965/

[54] M. Cloostermans, C. de Vos, T. Heida, A. de Keijzer, and M. van Putten, "Monitoring the brain in the adult ICU," in *Proceedings of the 4th Annual Symposium of the IEEE/EMBS Benelux Chapter*, Nov. 2009, pp. 128–130. [Online]. Available: http://doc.utwente.nl/70365/

[55] L. L. Leape, "Institute of medicine medical error figures are not exaggerated," *JAMA*, vol. 284, no. 1, pp. 95–97, 2000. [Online]. Available: http://dx.doi.org/10.1001/jama.284.1.95

[56] L. T. Kohn, J. M. Corrigan, and I. o. M. Molla S. Donaldson editors; Committee on Quality of Health Care in America, *To Err Is Human: Building a Safer Health System*. The National Academies Press, 2000. [Online]. Available: http://www.nap.edu/openbook.php?record_id=9728

[57] J. Dwyer, "Death of a boy prompts new medical efforts nationwide," *The New York Times*, October 25 2012. [Online]. Available: http://www.nytimes.com/2012/10/26/nyregion/tale-of-rory-stauntons-death-prompts-new-medical-efforts-nationwide.html

[58] J. Groopman, "Medical dispatches: What's the trouble? How doctors think." *The New Yorker*, Jan. 29 2007.

[59] P. Croskerry, "The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them," *Academic Medicine*, vol. 78, no. 8, pp. 775–780, Aug 2003.

[60] S. S. Lodder, "Computer assisted interpretation of the human eeg: improving diagnostic efficiency and consistency in clinical reviews," Ph.D. dissertation, Enschede, 2014. [Online]. Available: http://doc.utwente.nl/89312/

[61] M. Graber, R. Gordon, and N. Franklin, "Reducing diagnostic errors in medicine: what's the goal?" *Academic Medicine*, vol. 77, no. 10, pp. 981–992, 2002.

[62] S. G. Boodman, "Medical mystery: Alcoholism didn't cause man's diabetes and cirrhosis," *The Washington Post*, June 14 2011. [Online]. Available: http://www.washingtonpost.com/national/medical-mystery-alcoholism-didnt-cause-mans-diabetes-and-cirrhosis/2011/05/19/AGd0hdTH_story.html

[63] S. Gupta, T. Tran, W. Luo, D. Phung, R. L. Kennedy, A. Broad, D. Campbell, D. Kipp, M. Singh, M. Khasraw, L. Matheson, D. M. Ashley, and S. Venkatesh, "Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry," *BMJ Open*, vol. 4, no. 3, 2014. [Online]. Available: http://bmjopen.bmj.com/content/4/3/e004007.abstract

[64] G. Shafer, *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.

[65] L. A. Zadeh, "Book Review: A Mathematical Theory of Evidence," *AI Magazine*, vol. 5, no. 3, pp. 81–83, 1984.

[66] K. Sentz and S. Ferson, "Combination of evidence in Dempster-Shafer theory," Sandia National Laboratories, Tech. Rep. SAND2002-0835, 2002. [Online]. Available: http://prod.sandia.gov/techlib/access-control.cgi/2002/020835.pdf

[67] S. Salicone, *Measurement Uncertainty: An Approach via the Mathematical Theory of Evidence*, ser. Springer Series in Reliability Engineering. Springer US, 2007, ch. The Theory of Evidence, pp. 31–71.

[68] O. Užga-Rebrovs and G. Kuļešova, "A Comparative Analysis of Alternative Rules of Belief Combination," *Scientific Journal of RTU*, vol. 36, pp. 93–99, 2008.

[69] D. Nardia and R. J. Brachman, "An introduction to description logics," in *Description Logic Handbook*, F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, Eds. Cambridge University Press, 2002, pp. 5–44.

[70] R. M. Cooke and L. L. Goossens, "TU Delft expert judgment data base ," *Reliability Engineering & System Safety*, vol. 93, no. 5, pp. 657 – 674, 2008.

[71] F. Ouchi, "A literature review on the use of expert opinion in probabilistic risk analysis," The World Bank, Policy Research Working Paper Series 3201, Jan. 2004. [Online]. Available: http://ideas.repec.org/p/wbk/wbrwps/3201.html

[72] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, *Multiple Experts*. John Wiley & Sons, Ltd, 2006, pp. 179–192.

[73] J. Rougier, S. Sparks, L. Hill, and R. Sparks, *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge University Press, 2013.

[74] J. C. Hoffman and R. R. Murphy, "Comparison of bayesian and dempster-shafer theory for sensing: A practitioner's approach," in *in SPIE Proc. on Neural and Stochastic Methods in Image and Signal Processing II*, 1993, pp. 266–279.

[75] L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete, "Implementing relevance feedback in the bayesian network retrieval model," *Journal of the American Society for Information Science and Technology*, vol. 54, no. 4, pp. 302–313, Feb. 2003. [Online]. Available: http://dx.doi.org/10.1002/asi.10210

[76] J. Xin and J. S. Jin, "Relevance feedback for content-based image retrieval using bayesian network," in *Proceedings of the Pan-Sydney area workshop on Visual information processing*, ser. VIP '05. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2004, pp. 91–94. [Online]. Available: http://dl.acm.org/citation.cfm?id=1082121.1082137

[77] M. C. Florea, A.-L. Jousselme, ı. Bossé, and D. Grenier, "Robust combination rules for evidence theory," *Information Fusion*, vol. 10, no. 2, pp. 183–197, Apr. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.inffus.2008.08.007

[78] R. Haenni, "Shedding new light on Zadeh's criticism of Dempster's rule of combination," in *FUSION'05, 8th International Conference on Information Fusion*, vol. 2, Philadelphia, USA, pp. 879–884.

[79] O. Benjelloun, A. D. Sarma, A. Y. Halevy, and J. Widom, "ULDBs: Databases with Uncertainty and Lineage," in *VLDB*, U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, Eds. ACM, 2006, pp. 953–964.

[80] O. Benjelloun, A. Das Sarma, A. Halevy, M. Theobald, and J. Widom, "Databases with uncertainty and lineage," *The VLDB Journal*, vol. 17, pp. 243–264, 2008, 10.1007/s00778-007-0080-z. [Online]. Available: http://dx.doi.org/10.1007/s00778-007-0080-z

[81] J. Widom, *Managing and Mining Uncertain Data*, ser. Advances in Database Systems. Springer, 2009, vol. 35, ch. Trio: A System for Data, Uncertainty, and Lineage, pp. 113–148.

[82] L. Antova, C. Koch, and D. Olteanu, "$10^{(10^6)}$ worlds and beyond: Efficient representation and processing of incomplete information," *The VLDB Journal*, vol. 18, no. 5, pp. 1021–1040, Oct. 2009. [Online]. Available: http://dx.doi.org/10.1007/s00778-009-0149-y

[83] L. Antova, T. Jansen, C. Koch, and D. Olteanu, "Fast and simple relational processing of uncertain data," in *Proceedings of the 24th International IEEE Conference on Data Engineering (ICDE 2008)*. IEEE, 2008, pp. 983–992.

[84] M. Magnani and D. Montesi, "A survey on uncertainty management in data integration," *J. Data and Information Quality*, vol. 2, no. 1, pp. 5:1–5:33, Jul. 2010. [Online]. Available: http://doi.acm.org/10.1145/1805286.1805291

[85] M. van Keulen, "Managing uncertainty: The road towards better data interoperability," *IT - Information Technology*, vol. 54, no. 3, pp. 138–146, May 2012. [Online]. Available: http://dx.doi.org/10.1524/itit.2012.0674

[86] P. Agrawal and J. Widom, "Generalized Uncertain Databases: First Steps," in *MUD*, ser. CTIT Workshop Proceedings Series, A. de Keijzer and M. van Keulen, Eds., vol. WP10-04. Centre for Telematics and Information Technology (CTIT), University of Twente, The Netherlands, 2010, pp. 99–111.

[87] A. de Keijzer and M. van Keulen, "User Feedback in Probabilistic Integration," in *Second International Workshop on Flexible Database and Information System Technology (FlexDBIST 2007), Regensburg, Germany*. Los Alamitos: IEEE Computer Society Press, September 2007, pp. 377–381.

[88] M. van Keulen and A. de Keijzer, "Qualitative Effects of Knowledge Rules and User Feedback in Probabilistic Data Integration," *The VLDB Journal*, vol. 18, no. 5, pp. 1191–1217, October 2009.

[89] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," *Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on*, vol. 48, no. 2, pp. 177–183, 2001.

[90] F. S. Bao, X. Liu, and C. Zhang, "Pyeeg: an open source python module for eeg/meg feature extraction," *Computational intelligence and neuroscience*, vol. 2011, 2011.

[91] G. Lin and L. Chen, "A Grid and Fractal Dimension-Based Data Stream Clustering Algorithm," *Information Science and Engieering, International Symposium on*, vol. 1, pp. 66–70, 2008.

[92] D. Barbará and P. Chen, "Using the fractal dimension to cluster datasets," in *KDD*, 2000, pp. 260–264.

[93] G. Yan and Z. Li, "Using cluster similarity to detect natural cluster hierarchies," in *FSKD (2)*, 2007, pp. 291–295.

[94] M. Sarkar and T.-Y. Leong, "Characterization of medical time series using fuzzy similarity-based fractal dimensions," *Artificial Intelligence in Medicine*, vol. 27, no. 2, pp. 201–222, 2003.

[95] A. Eke, P. Herman, L. Kocsis, and L. Kozak, "Fractal characterization of complexity in temporal physiological signals," *Physiological measurement*, vol. 23, no. 1, pp. R–R38, 2002.

[96] A. Accardo, M. Affinito, M. Carrozzi, and F. Bouquet, "Use of the fractal dimension for the analysis of electroencephalographic time series." *Biological Cybernetics*, vol. 77, no. 5, pp. 339–350, 1997.

[97] Mehmet Malcok and Y. Alp Aslandogan and Ayin Yesildirek, "Fractal dimension and similarity search in high-dimensional spatial databases," in *IRI*, 2006, pp. 380–384.

[98] D. S. Mazel and M. H. Hayes, "Fractal modeling of time-series data," in *Conference Record of the Twenty-Third Asilomar Conference of Signals, Systems and Computers*, 1989, pp. 182–186.

[99] M. Barnsley, *Fractals everywhere*, 2nd ed. San Diego, CA, USA: Academic Press Professional, Inc., 1988.

[100] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal eeg patterns," in *Computer-Based Medical Systems, 1995., Proceedings of the Eighth IEEE Symposium on*. IEEE, 1995, pp. 212–217.

[101] K. Kalpakis, D. Gada, and V. Puttagunta, "Distance Measures for Effective Clustering of ARIMA Time-Series," in *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2001, pp. 273–280.

[102] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[103] C. M. Bishop, M. Svensén, and C. K. Williams, "Em optimization of latent-variable density models," in *Advances in Neural Information Processing Systems*, 1996, pp. 465–471.

[104] F. D. College and F. Dellaert, "The expectation maximization algorithm," Tech. Rep., 2002.

[105] C. M. Bishop, "Pattern recognition and machine learning," 2006.

[106] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[107] M. R. Gupta and Y. Chen, "Theory and use of the em algorithm," *Foundations and Trends in Signal Processing*, vol. 4, no. 3, pp. 223–296, 2011. [Online]. Available: http://dx.doi.org/10.1561/2000000034

[108] M. De Hoon, S. Imoto, J. Nolan, and S. Miyano, "Open source clustering software," *Bioinformatics*, vol. 20, pp. 1453–1454, June 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1092875.1092876

[109] A. Climescu-Haulica, "How to Choose the Number of Clusters: The Cramer Multiplicity Solution," in *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V.*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, R. Decker and H.-J. Lenz, Eds. Freie Universität Berlin: Springer, March 8-10 2006, pp. 15–22.

[110] E. Avramidis, "Rankeval: Open tool for evaluation of machine-learned ranking," *The Prague Bulletin of Mathematical Linguistics*, vol. 100, pp. 63–72, 2013.

[111] R. Korra, P. Sujatha, P. Dhavachelvan, M. N. Kumar, and S. Chetana, "Performance assessment of amrr and adcg metrics in mlir and ir systems," *International Journal of Computer Applications*, vol. 24, no. 2, 2011.

[112] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.

[113] Q. Shen, Y. Yang, Z. Wu, X. Yang, L. Zhang, X. Yu, Z. Lao, D. Wang, and M. Long, "Sapsc: Security architecture of private storage cloud based on hdfs," in *Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on*, March 2012, pp. 1292–1297.

[114] H. Jing, L. Renfa, and T. Zhuo, "The research of the data security for cloud disk based on the hadoop framework," in *Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on*, June 2013, pp. 293–298.

[115] G. H. Limited, "http://www.medcyclopaedia.com/library/," medical Encyclopedia. [Online]. Available: http://www.medcyclopaedia.com/library/

[116] E. H. Chudler, http://faculty.washington.edu/chudler/1020.html, 1996-2009. [Online]. Available: http://faculty.washington.edu/chudler/1020.html

About the author

Ghita Berrada was born in 2 January 1984 in Rabat, Morocco.

She obtained an MSc by research in Pattern Analysis and Neural Networks from Aston University (Birmingham, UK) as well as an engineering degree (equivalent to an MSc, with a major in Computer Science) from École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE) (Evry, France) in 2008. She started her PhD at the University of Twente, The Netherlands in September 2009.

The goal of her PhD project was to design a medical data sharing platform, using EEG data as example of medical data, in order to support the diagnosis process.

Her research interests include data mining, machine learning, databases, uncertain data management and similarity search.