Knowledge discovery in databases using novel approaches based on Computational Intelligence

Zaineb Chelly Dagdia

Habilitation à diriger des recherches

université PARIS-SACLAY

Contributions en Extraction de Connaissances en utilisant des Nouvelles Approches basées sur l'Intelligence Computationnelle

Habilitation à diriger des recherches de l'Université Paris-Saclay

présentée et soutenue à l'Université de Versailles Saint-Quentin-en-Yvelines, UFR des Sciences, Paris-Saclay, le 24 Octobre 2023, par

Zaineb CHELLY DAGDIA Ep. GARCIA

Composition du jury

Dominik Ślęzak Professeur, Université de Varsovie	Rapporteur
Guillaume Cleuziou Professeur, Université d'Orléans	Rapporteur
Pierre Gançarski Professeur, Université de Strasbourg	Rapporteur
Sihem Amer-Yahia Directrice de recherche, Laboratoire d'Informatique de Grenoble	Examinatrice
Engelbert Mephu Nguifo Professeur, Université Blaise Pascal Clermont-Ferrand	Examinateur
Fariza Tahi Professeur, Université d'Evry	Présidente

université
PARIS-SACLAY

Titre : Contributions en Extraction de Connaissances en utilisant des Nouvelles Approches basées sur l'Intelligence Computationnelle

Mots clés : Extraction de connaissances, Intelligence computationnelle, Approches axées sur les données

Résumé : La thèse d'habilitation explore différentes méthodes visant à découvrir des connaissances à partir des données en utilisant de nouvelles techniques basées sur l'intelligence computationnelle. Ces méthodes ont été développées pour surmonter les divers défis associés à l'extraction de connaissances, tels que le prétraitement des données, l'environnement incertain, l'optimisation, la scalabilité et le respect des données privées. De plus, la thèse présente des adaptations d'approches computationnelles visant à améliorer des domaines spécifiques tels que l'exploration et l'analyse des données de mobilité et la cybersécurité. Ces adaptations mettent en évidence le potentiel des techniques d'intelligence computationnelle pour relever efficacement des défis concrets.

La thèse explore également des approches de soins de santé personnalisés, démontrant comment les techniques avancées d'analyse de données et d'apprentissage automatique peuvent être utilisées pour améliorer les résultats individuels des patients. Cela souligne l'importance d'exploiter les méthodes basées sur les données pour élaborer des plans de traitement personnalisés, adaptés aux besoins et aux circonstances uniques de chaque patient.

Title: Knowledge discovery in databases using novel approaches based on Computational Intelligence

Keywords: Knowledge extraction, Computational intelligence, Data-driven approaches

Abstract: The Habilitation Thesis explores various methods aimed at discovering knowledge from data using new techniques based on computational intelligence. These methods have been developed to overcome diverse challenges associated with knowledge extraction, such as data preprocessing, uncertain environments, optimization, scalability, and privacy preservation. Additionally, the thesis presents adaptations of computational approaches aimed at improving specific domains such as mobility data mining and cybersecurity. These adaptations highlight the potential of computational intelligence techniques to effectively address practical challenges.

Furthermore, the thesis explores personalized healthcare approaches, demonstrating how advanced data analysis and machine learning techniques can be used to enhance individual patient outcomes. This underscores the importance of leveraging data-driven methods to develop personalized treatment plans tailored to the unique needs and circumstances of each patient.

Acknowledgments

First and foremost, I express my gratitude and respect to Mr. Dominik Ślęzak (Professor at the University of Warsaw), Mr. Guillaume Cleuziou (Professor at the University of Orléans), and Mr. Pierre Gançarski (Professor at the University of Strasbourg) for the honor of accepting to serve as rapporteurs for this habilitation thesis. I extend my sincerest appreciation to Ms. Sihem Amer-Yahia (Research Director, CNRS, Grenoble Computer Science Laboratory), Mr. Engelbert Mephu Nguifo (Professor at Blaise Pascal University, Clermont-Ferrand), and Ms. Fariza Tahi (Professor at the University of Evry) for graciously accepting the role of jury members. I also thank my mentor Prof Karine Zeitouni for her guidance and support throughout this habilitation thesis journey.

The realization of this work owes a debt of gratitude to the support from individuals who graciously welcomed me into their teams and laboratories across various esteemed organizations such as The National Institute for Research in Digital Science and Technology (Inria), Aberystwyth University, Granada University, and Assistance Publique - Hôpitaux de Paris (APHP). I am deeply appreciative of the fruitful exchanges, scientific collaboration, and continuous support that I have received from these remarkable individuals. Throughout the course of this endeavor, I had the privilege of collaborating on several projects with talented doctoral students, and their dedication and contributions have brought me immense satisfaction. I extend my heartfelt gratitude to all the members of the Ambient Data Access and Mining team and the Data and Algorithms for an Intelligent and Sustainable City Laboratory for their support. I would like to express my thanks to all those who have directly or indirectly contributed to this work. Furthermore, I would like to acknowledge my colleagues at UVSQ.

I want to extend special thanks to my immediate and extended family, who have been a constant source of support and inspiration throughout my journey. In particular, I am deeply grateful to my loving husband, Stephane, whose unwavering belief in me has been a pillar of strength. To my cherished daughter, Julia, who fills my life with joy and purpose, I am eternally grateful. Additionally, I would like to acknowledge my parents, sisters, and brother, whose love and encouragement have been a constant source of motivation. Words cannot fully convey the depth of my appreciation, as they are well aware that my accomplishments would not be possible without their support and belief in me.

Contents

List of Figures

List of Tables

Introduction

In the dynamic landscape of data-driven research and analysis, the fields of data mining and knowledge discovery in databases have emerged as focal points of immense interest. As we navigate the vast expanse of the big data era, where information is generated at an unprecedented rate, the significance of efficiently preprocessing this data becomes paramount. However, the challenges in this domain extend beyond data volume and complexity. In addition to addressing the intricacies of big data settings, researchers must confront a myriad of other obstacles, including uncertain environments, optimization challenges, and the practical application of their findings in real-world scenarios. Moreover, as privacy and data protection concerns continue to intensify, it becomes increasingly imperative to develop robust methods that not only harness the power of big data but also safeguard individual privacy rights. In this multifaceted landscape, my research aims to explore, innovate, and forge new pathways, unraveling the intricacies of data preprocessing, addressing uncertainties, optimizing performance, dealing with privacy issues, and uncovering practical applications that can propel data-driven research forward.

The conducted research that will be described in this Habilitation Thesis is structured around the following three main research directions:

- **Research Direction 1:** *New approaches for knowledge discovery and privacy preservation using granular computation and federated learning.* In the first aspect of my research, I delve into the exploration and addressing of two significant challenges: big data preprocessing, with a special focus on feature selection, and privacy concerns. To tackle these challenges, I develop innovative methods that leverage the power of granular computation and federated learning. Specifically, my research presents new approaches for preprocessing big data by utilizing granular computation, highlighting its effectiveness in handling large datasets. Furthermore, my investigation extends to the mining of semantically enriched mobility data, considering both the perspective of granular computation and the principles of federated learning, with an emphasis on privacy-conscious analysis and preservation.
- **Research Direction 2:** *New approaches for knowledge discovery using biological computation and deep learning.* In the second research domain, my focus lies on the study of artificial immune systems, evolutionary computation, and deep learning. Diving deep into this investigation, I thoroughly analyze the intrinsic characteristics and capabilities of these computational intelligence based approaches to effectively address complex problems in diverse scenarios. Uncertain environments, optimization

challenges, big data settings, and real-world applications are among the scenarios I explore. The main objective of this exploration is to enhance and adapt the computational intelligence based techniques to suit a range of various applications, including wireless sensor network deployment, malware detection, and the context of Bring Your Own Device (BYOD). By thoroughly examining the strengths and capabilities of these computational intelligence based techniques, my research aims to unlock their potential and pave the way for innovative solutions in various domains.

- **Research Direction 3:** *Personalizing healthcare: Focus on real-world applications.* The third area of research is dedicated to the personalization of healthcare in real-world settings, with a strong emphasis on collaboration with hospitals. This research primarily focuses on case studies in (i) epidemiology for cancer incidence prediction, and (iii) personalized approaches to the treatment of sepsis.

These research areas have been explored in collaboration with numerous partners from academia and industry as well as Master, PhD students, and postgraduates.

The Habilitation Thesis will present an overview of the main ideas and key findings of each contribution in these three research areas. Detailed technical information about each contribution can be found in the relevant published research papers.

Habilitation Thesis outline

This Habilitation Thesis is structured into two parts, consisting of seven main chapters.

Part I: Activity reports

- **Chapter 1:** *My professional and academic journey.* This chapter reflects on the path I have traversed in my professional and academic pursuits.
- **Chapter 2:** *In the realm of research and the development of innovation and technology transfer.* The chapter highlights various projects, publications, acknowledgments, engagements, and responsibilities related to my research activities and to the development of innovation and technology transfer.
- **Chapter 3:** *Aptitude for supervision.* In this chapter, I emphasize my involvement in supervising and mentoring students across different academic levels.

Part II: Summary of scientific contributions

- **Chapter 4:** *New approaches for knowledge discovery and privacy preservation using granular computation and federated learning.* This chapter presents the first research direction. It introduces innovative approaches for preprocessing big data through the utilization of rough sets as a granular computation technique. Additionally, it explores the use of granular computation and federated learning

as means to address privacy preservation concerns.

- **Chapter 5:** *New approaches for knowledge discovery using biological computation and deep learning.* This chapter presents the second research direction. The chapter outlines various contributions in the fields of artificial immune systems, evolutionary computation, and deep learning. It encompasses both theoretical work and practical adaptations of these approaches across diverse application domains.
- **Chapter 6:** *Personalizing healthcare: Focus on real-world applications.* This chapter, which presents the third research direction, takes some of the conducted research presented in this Habilitation Thesis into practice.
- **Chapter 7:** *List of publications.* In this chapter, a list of my published works is presented.

Finally, this Habilitation Thesis ends with a conclusion and two Appendices. The conclusion gives a summary of the presented work and offers perspectives for future research. **Appendix 8** presents the main concepts of the dendritic cell algorithm, and **Appendix 9** presents the main concepts of rough set theory, as a granular computation approach, for feature selection.

Part I

Activity reports

Chapter 1

My professional and academic journey

1.1 Introduction

In this chapter, I delve into a retrospective view of my professional and academic path. This chapter is structured as follows: Section 1.2 provides insight into my professional experiences, while Section 1.3 offers an overview of my academic progression. Section 1.4 delves into the challenges tackled in my Habilitation research including a brief description of my research directions and methodology. The conclusion is given in Section 1.5.

1.2 A glint about my professional journey

In this section, I will reflect on my professional journey. I will provide an overview of the commitments I made and responsibilities I took in each position. Table 1.1 gives a clear and concise picture of my professional path.

Table 1.1: Summary of my professional path

Period	**Professional title**	**University/Institute/Faculty**
2020 -	**Associate Professor**	University of Versailles Saint-Quentin-en-Yvelines (UVSQ), UFR des Sciences, Paris-Saclay ***Laboratory:*** Data and Algorithms for an Intelligent and Sustainable City (DAVID) Laboratory ***Research group:*** Ambient Data Access and Mining (ADAM)

2019 - 2020	**Research & Development (R&D) Technical Project Manager** *(50%)* and **Partnership and Innovation Project Lead** *(50%)* ********** **R&D Technical Project Manager** **Partnership and Innovation Project Lead**	The National Institute for Research in Digital Science and Technology (Inria), Nancy ***Research group:*** MULTISPEECH ***Department:*** Technology Transfer, Innovation and Partnership
2017 - 2019	**Marie SkŁodowska Curie Research Fellow**	Aberystwyth University, UK ***Research group:*** Advanced Reasoning (ARG)
2016 - 2017	**Assistant Professor**	High Institute of Management of Tunis (ISG), Tunis, Tunisia ***Laboratory:*** LAboratory of Operation Researches, DEcision and process Control (LARODEC)
2013 - 2016	**Lecturer**	High Institute of Management of Tunis (ISG), Tunis, Tunisia ***Laboratory:*** LAboratory of Operation Researches, DEcision and process Control (LARODEC)
2010 - 2013	**Teaching Assistant**	Faculty of Economics and Management of Nabeul (FSEGN), Nabeul, Tunisia

		Laboratory: LAboratory of Operation Researches, DEcision and process Control (LARODEC)

1.2.1 Overview of teaching activities

FSEGN and ISG, Tunisia (2010 - 2017)

My career in higher education began in September 2010 at FSEGN where I was hired as a ***Teaching Assistant*** in Computer Science at the Department of Quantitative Methods and Computer Science. I held this position for three consecutive academic years from 2010 to 2013. In June 2013, I was successful in the national exam for the position of ***Lecturer*** and was transferred to ISG within the Department of Computer Science, where I held this position from 2013 to 2016. In 2016, I was promoted to the position of ***Assistant Professor*** at ISG. However, I only taught for one semester during the academic year as I was granted a detachment in February 2017 to conduct my research project in the United Kingdom where I held the position of a Marie SkŁodowska Curie Research Fellow.

Over the period of 2010 to 2017, I accumulated over **1 650 hours** of teaching experience at both the FSEGN and ISG, covering a wide range of subjects taught to students of various academic levels. This included both lectures and hands-on practical sessions. In addition to my teaching duties, I was also responsible for supervising students at different stages of their academic careers. This is described in more detail in Chapter 3. Furthermore, I had various pedagogical, team-oriented, and administrative responsibilities, among these I mention the following:

- Course development
- Designing assignments and projects
- Assessment
- Analyzing student data
- Serving on academic committees
- External partnerships
- Engaging in outreach

Further institutional responsibilities are outlined in Chapter 2 – Section 2.6. Figures 1.1, 1.2, and 1.3 show the full list of lectures and practical sessions I gave at FSEGN and ISG. Please, note that the names of modules are kept in French. In the figures, LAIG, LFIG and LAID refer to Licence Appliquée en Informatique de Gestion, Licence Fondamentale en Informatique de Gestion, and to Licence Appliquée en Informatique Décisionnelle, respectively.

Année universitaire	Module	Niveau	Nature	Volume horaire hebdomadaire
2010-2011 2011-2012	Logiciel Libre **(semestre 1)**	L1 LAIG & L1 LFIG	Cours magistral	3h
	Algorithmique et Structures de Données I **(semestre 1)**	L1 LAIG	TD	6h
	E-Learning **(semestre 2)**	L1 LAIG & L1 LFIG	Cours magistral	3h
	E-commerce **(semestre 1)**	L3 LFIG	Cours magistral	1.5
	EDI et commerce électronique **(semestre 1)**	L1 LAIG	TD	1.5
	Algorithmique et Structures de Données II **(semestre 2)**	L1 LFIG	TD	6h
2012-2013	Algorithmique et Structures de Données I **(semestre 1)**	L1 LAIG	TD	12h
	Algorithmique et Structures de Données II **(semestre 2)**	L1 LFIG	TD	12h

Figure 1.1: Teaching load – FSEGN

Année universitaire	Module	Niveau	Nature	Volume horaire hebdomadaire
2013-2014	Algorithmique et Structures de Données I **(semestre 1)**	L1 LFIG	TD	3h
	Atelier de programmation I **(semestre 1)**		Cours	6h
	Algorithmique et Structures de Données II **(semestre 2)**		TD	6h
2014-2015	Algorithmique et Structures de Données I **(semestre 1)**		TD	6h
	Atelier de programmation I **(semestre 1)**		Cours	4h30
	Algorithmique et Structures de Données II **(semestre 2)**		TD	6h
	Atelier de programmation II **(semestre 2)**		TD	3h
2015-2016	Algorithmique et Structures de Données I **(semestre 1)**		TD	3h
	Atelier de programmation I **(semestre 1)**		Cours	1h30
	Architecture et Systèmes **(semestre 1)**	L1 LAID	Cours	3h
	Algorithmique et Structures de Données II **(semestre 2)**	L1 LFIG	TD	6h
2016-2017	Algorithmique et Structures de Données I **(semestre 1)**	L1 LFIG	TD	3h
	Atelier de programmation I **(semestre 1)**		Cours	4h30

Figure 1.2: Teaching load – ISG

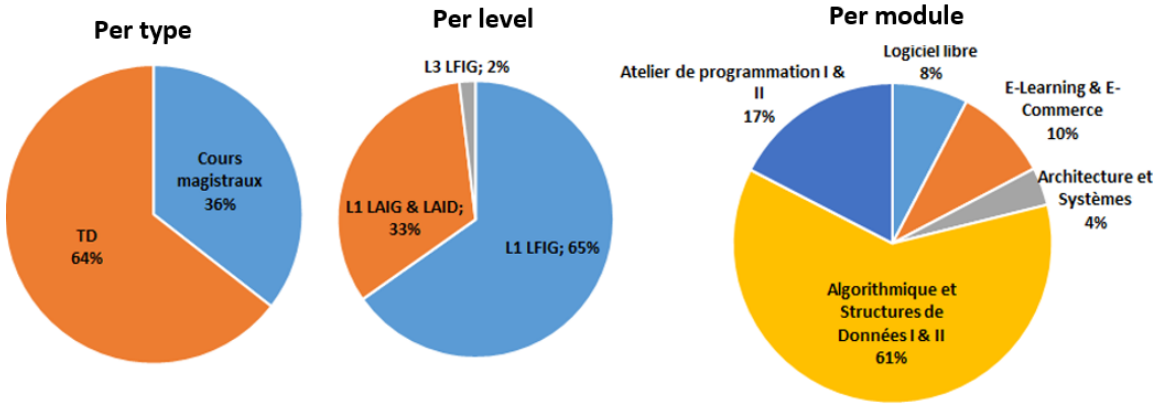

Figure 1.3: Summary of teaching by type, level, and module

Training programs and certificates. I have also participated in several training programs organized by the *"Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Tunisie"* and the *"Centre de Didactique et de la Pédagogie Universitaires (CDPU)"*, among

these I mention the list below.

- ❒ "Analyse des pratiques professionnelles" (2015)
- ❒ "Communication pédagogique" (2015)
- ❒ "Relations pédagogique" (2015)
- ❒ "Planification d'un cours" (2014)
- ❒ "Méthodes et Techniques d'enseignement" (2014)
- ❒ "Évaluation des acquis des étudiants" (2014)

I have obtained the Teaching Certificate "University Didactic and Pedagogy" – **ISO 29990 certificate** (2014): *Services de formation dans le cadre de l'éducation et de la formation non formelles – Exigences de base pour les prestataires de services.*

UVSQ, Paris-Saclay (2020 - current)

In 2020, without any opened permanent position at Inria in my fields of interests, I decided to apply for an ***Associate Professor*** position at UVSQ, Paris-Saclay and got it at the first attempt. I joined the *DAVID Laboratory*; and specifically the *ADAM* research group.

At UVSQ, my teaching duties have been varied, covering a range of subjects for students of all levels, from bachelor to master level. Since I joined UVSQ in 2020, I have taught a total of **627.5 hours (HETD)**. Figure 1.4 shows the full list of lectures and practical sessions I gave at UVSQ, Paris-Saclay.

Training programs and certificates. I have also taken on a number of pedagogical, team-oriented, and administrative responsibilities; some of which are mentioned among the responsibilities taken at FSEGN and ISG. I have also participated in several training programs, among these I mention the following:

- "Formation à l'encadrement des doctorant(e)s" organized by Paris-Saclay (2022)
- "Alignement pédagogique" organized by Paris-Saclay (2020)
- "Adapter son enseignement au distanciel" organized by Paris-Saclay (2020)
- "Comment tirer un bilan collectif d'une année de pédagogie distancielle ? Journée Fil Rouge 2021 L'évaluation en questions" organized by Paris-Saclay (2020)
- "Enseigner et former dans l'enseignement supérieur" organized by MOOC (2020)
- "La classe inversée à l'ère du numérique" organized by FUN MOOC (2020)
- "S'initier à l'enseignement en Sciences Numériques et Technologie" organized by FUN MOOC (2020)

- "Utilisation des espaces physiques en situation hybride" organized by UNIF (2020)
- "Le mind mapping" organized by UNIF (2020)
- "Evaluation en ligne" organized by UNIF (2020)
- "Comment dynamiser ses cours à distance !" organized by Sorbonne Paris Nord (2020)

1.2.2 Overview of research and leadership activities

LARODEC, Tunisia (2008 - 2017)

My passion to research has begun since 2007 - 2008 when I was taking my final year as a Bachelor student at FSEGN, Tunisia. For my final year project, and in the context of a real case study which was made possible via two secondments, I developed a *"Fuzzy Quality Audit Expert System" (SEF-AQ)* that automates the process of quality audits. The system aimed to verify if a company's quality process complies with the requirements of ISO 9001 version 2000. After my graduation, I was eager to continue the research path I initially chose. To fulfill this vision, I took my Master degree at ISG, Tunisia, where I proposed a new algorithm in the theme of Artificial Immune Systems (AIS), named *"A Fuzzy Dendritic Cell Method" (FDCM)*. For my PhD thesis, I continued my research in the field of AIS by developing novel learning models inspired by the behavior of dendritic cells. A comprehensive overview of each of these research developments will be provided in Section 1.3.

ARG (MSCA), United Kingdom (2017 - 2019)

After earning my PhD in 2014, I was eager to start my own research project and applied for the prestigious ***Marie Skłodowska-Curie Actions (MSCA) – Individual Fellowship*** under the *Horizon 2020 European Research and Innovation programme*'s Excellent Science pillar. My project, entitled *"Optimised Framework based on Rough Set Theory for Big Data Pre-processing in Certain and Imprecise Contexts"* was accepted by the European Commission based on a budget allocation of 183.5 K€, in early 2016, and I officially began working on it in March 2017 as a Marie Skłodowska-Curie research fellow at Aberystwyth University, UK. My 2-years research project (2017 - 2019) was carried out in collaboration with various academic and non-academic partners, including Sorbonne Paris Nord University (France), University of Granada (Spain), University of Birmingham (United Kingdom), and Hôtel-Dieu Hospital in Paris (France). The project is described in Chapter 2, Table 2.8.

My main tasks in this project were (i) the development of the project's different work packages, (ii) financial monitoring of the project in collaboration with the financial department of Aberystwyth University, (iii) disseminating research results on an international scale, (iv) leading communications and interacting with the European Commission for project-related reporting tasks, (v) ensuring local integration within the university (research team, department, etc.) and national (UK) and international outreach, (vi) providing tutoring and

Année	Niveau	Diplôme	Intitulé	Type de formation	Nature	Effectifs	Volume horaire annuel (h)
2022-2023	1ère année	Licence	Fondements de l'Informatique I **(2 CM, 14 TD)**	Formation initiale, présentielle	Cours magistral	72	18
					TD	62	72
	2ème année	Licence	Systèmes d'Exploitation **(1 CM, 4 TD)**	Formation initiale, présentielle	TD	33	36
	2ème année	M2 Datascale	Projet conception	Formation initiale, présentielle	Encadrement pour la réalisation d'un projet de conception orienté recherche	17	25.5
	2ème année	M2 Datascale	Projet programmation	Formation initiale, présentielle	Encadrement pour la réalisation d'un projet de programmation orienté recherche	17	25.5
	2ème année	M2 Datascale	Fouille de données **(1 CM, 1 TD)** **Responsable de l'UE**	Formation initiale, présentielle	Cours magistral	30	15
					TD	30	6
						Total charge :	**214.5**
						Heures complémentaires :	**22.5**
2021-2022	**Réduction de service MCF – Congé Maternité**						**96**
	2ème année	Licence	Systèmes d'Exploitation **(1 CM, 4 TD)**	Formation initiale, présentielle	TD	32	72
	2ème année	M2 Datascale	Projet conception	Formation initiale, présentielle	Encadrement pour la réalisation d'un projet de conception orienté recherche	5	7.5
	2ème année	M2 Datascale	Projet programmation	Formation initiale, présentielle	Encadrement pour la réalisation d'un projet de programmation orienté recherche	5	7.5
	2ème année	M2 Datascale	Fouille de données **(1 CM, 1 TD)** **Responsable de l'UE**	Formation initiale, présentielle	Cours magistral	36	12
					TD	36	3
						Total charge :	**204**
						Heures complémentaires :	**12**
2020-2021	**Réduction de service MCF**						**32**
	1ère année	Licence	Fondements de l'Informatique I **(5 CM, 19 TD)**	Formation initiale, présentielle puis à distance (en raison de la situation sanitaire COVID)	Cours magistral	51	18
					TD	86	108
	1ère année	Licence	Fondements de l'Informatique II **(2 CM, 13 TD)**	Formation initiale, présentielle puis à distance (en raison de la situation sanitaire COVID)	TD	20	6
	2ème année	Licence	Systèmes d'Exploitation **(1 CM, 4 TD)**	Formation initiale, présentielle puis à distance (en raison de la situation sanitaire COVID)	TD	28	36
						Total charge :	**209**
						Heures complémentaires :	**17**

Figure 1.4: Teaching load – UVSQ

seminars related to the project, (vii) creating opportunities for setting up new European projects, and (viii) establishing new research collaborations.

In addition, I took roles and responsibilities as an **Ambassador** of the MSC Individual Fellowship program. My main responsibilities will be discussed in Chapter 2 – Section 2.3.1.

Inria (2019 - 2020)

After successfully completing my MSC Individual Fellowship, I sought to further expand my expertise. I was offered the opportunity to take on a dual job at the *Inria*, Nancy. As a ***R&D Technical Project Manager*** for a H2020 ICT project, I was responsible for overseeing the development and execution of the project, utilizing 50% of my working time. For the remaining half of my time, I served as a ***Partnership and Innovation Project Lead***, where I was responsible for establishing and maintaining strategic partnerships, as well as identifying and developing new opportunities for innovation. This role allowed me to broaden my skill set and deepen my expertise in these key areas.

My main responsibilities within the MULTISPEECH research team will be described in Chapter 2, Section 2.4.1. Additionally, as a member of the MULTISPEECH research team and as a researcher, I worked on the following research themes: "big data pre-processing methods", "intrusion detection", and "evolutionary algorithms and optimization". My responsibilities at the Technology Transfer, Innovation and Partnership department at Inria will be described in Chapter 2, Section 2.5.

ADAM/DAVID, UVSQ, Paris-Saclay (2020 - current)

During my tenure at UVSQ, I carried forward my ongoing research endeavors from before joining the university, while also exploring new topics in close collaboration with fellow members of the ADAM team. Additionally, I fostered partnerships with entities outside of academia, such as Assistance publique – Hôpitaux de Paris (APHP). Among the research activities I have undertaken since joining UVSQ in 2020, I can provide a non-exhaustive list of the following contributions:

- ***Research projects*** (details can be found in Chapter 2):
  - Ongoing projects: I am currently involved in 5 active projects, as a work-package co-leader (4 projects) and as an involved researcher (1 project).
- ***Co-supervision*** (details can be found in Chapter 3):
  - 3 PhD students (since 2022)
  - 1 PostDoc (2022)
  - 1 Software Engineer (2020 - 2021)
- I am also actively engaged in several other research activities. These activities include organizing scientific events (e.g., workshops, special sessions), presenting as an invited speaker, and serving as a program committee member and reviewer for international

conferences and reputable journals. The specifics of these activities can be found in more details in Chapter 2.

1.3 A glint about my academic journey

A summary of my obtained diplomas is presented in Table 1.2. The details of my Master and PhD research work are provided in what follows.

Table 1.2: Summary of the obtained diplomas

Diploma	Enrollment	Defense	Distinction	Institute/Faculty
PhD	2011 - 2014	2014	Very honorable	ISG
Master	2008 - 2010	2010	Very good	ISG
Bachelor	2004 - 2008	2008	Good	FSEGN

1.3.1 Master Thesis

Institute:	ISG, Tunisia
Laboratory:	LARODEC
MSc Thesis title:	Handling Imprecision in Danger Theory using Fuzzy Set Theory: A Fuzzy Dendritic Cell Method (FDCM)
Viva:	09 June 2010 — obtained with distinction
Committee Members:	**Chair:** Prof. Rim Faiz, Carthage High Commercial Studies Institute, Tunisia
	Reviewer: Prof. Nadia Soussi, ISG
	Advisor: Prof. Zied Elouedi, ISG

Context

Danger Theory (DT) [Mat01], a widely accepted concept among immunologists, suggests that the immune system not only distinguishes between self and non-self, but also identifies and responds to potentially harmful elements and events. To detect these danger-producing elements, the immunological theory is based on the behaviour of special natural immune cells called the Dendritic Cells (DCs). An inspiration from the DCs behavior led to the development of an immune classification algorithm called the Dendritic Cell Algorithm (DCA) [GAT06] (see Appendix 8). DCA was successfully applied to a wide range of real-world

applications [CE16b]. This is due to the worthy characteristics expressed by the algorithm as it exhibits several interesting and potentially beneficial features for binary classification problems. This caught the attention of many researchers to explore more the algorithm and to investigate its behaviour.

It is in this emerging field of research, that my Master Thesis was subscribed. My Thesis focused on exploring the behaviour of the DCA to discern its characteristics and performance as a binary classification algorithm. It was essential to study the functioning of the DCA and to highlight its advantages as well as to propose adequate solutions to overcome its encountered and observed limitations. This is to increase the accessibility of this algorithm to future users.

Problem statement

After a thorough study of the DCA, it was observed that the application of this binary algorithm is limited to a specific order of the classes, i.e., sensitive to the input class data order. Concretely, when the algorithm is applied to databases having ordered data items between the two possible classes, that is to say where all the instances of the "normal" class are directly followed by all instances belonging to the "anomaly" class, the DCA generates interesting classification results. Otherwise, the results of the DCA are unsatisfactory. When going through the DCA algorithmic details, and specifically the step that promotes migrated cells to either the semi-mature (safe) or to the mature (danger) context depending on their accumulated response — called the context assessment phase —, I have noticed that this phase relies on a crisp boundary. In other words, in order to assign a specific context (semi-mature or mature) to each data item, it is necessary to apply a crisp comparison between the values of these two contexts. Nevertheless, such strict comparison ignores cases where the difference value between the two context values is small, resulting in a case where the final context of the object is hard to be defined. Another issue I noticed with the standard DCA is the imprecision found in the definition of some words such as "semi-mature" and "mature" which are quantified numerically. It is considered as unrealistic to assign a precise value to the cell context since it is difficult to fix the range or extents of the semantic of a semi-mature or a mature context. To overcome these limitations, I proposed a Fuzzy Dendritic Cell Method (FDCM) which is based on the fundamental concepts of Fuzzy Set Theory.

Contributions and publications

The developed Fuzzy Dendritic Cell Method (FDCM) [CE10] is based on the same classical DCA algorithmic steps (see Appendix 8) except for the context assessment phase. The new version of this phase is based on the concepts of fuzzy set theory to allow the algorithm to handle the crisp environment as well as the encountered imprecision. The fuzzy context assessment phase involves the following tasks:

1. Describing the context of each data item using linguistic variables.
2. Building a knowledge base comprising rules to support the fuzzy inference. The different rules are extracted from the information reflecting the effect of each input signal on the state of an immunological dendritic cell.

Our results in [CE10] showed that FDCM witnessed a significant improvement in classification accuracy, unlike the DCA, in the case of a high rate of class randomization.

The findings of this work were published in the 10th International Conference of Artificial Immune Systems, Springer, ICARIS'2011[1] [CE10].

1.3.2 PhD Thesis

Institute:	ISG, Tunisia
Laboratory:	LARODEC
Thesis Title:	New Danger Classification Methods in an Imprecise Framework
Viva:	28 August 2014 — obtained with honorable distinction
Committee Members:	**Chair:** Prof. Nahla Ben Amor, ISG
	Reviewer: Prof. Talel Ladhari, High School of Economic and Commercial Sciences of Tunis, Tunisia
	Reviewer: Dr. Salah Ben Abdallah, Tunis Business School, Tunisia
	Member: Dr. Lamia Labed Jilani, ISG
	Advisor: Prof. Zied Elouedi, ISG

Context

As mentioned, the DCA is a new foundation in the field of artificial immune systems. As a new paradigm, more work has to be done on the algorithm in order to study its properties and its performance as a binary classifier. In this context, the main objective of my PhD Thesis was to investigate a number of algorithmic properties of the DCA.

[1] ICARIS joined the Genetic and Evolutionary Computation Conference (GECCO) (ranked A) since 2013.

Problem statement

I performed a detailed study of the DCA behaviour. While doing so, I noticed that the algorithm has two critical limitations. The first limitation is linked to the DCA data pre-processing phase while the second limitation is related to the algorithm sensitivity to the input class data order. More precisely, these two shortcomings are explained as follows:

- *Data pre-processing:* While investigating the DCA data pre-processing phase, I noticed that this phase is not robust as it is based on the use of feature reduction techniques. In fact, applying feature reduction techniques destroys the underling meaning behind the features present in the used data set. Losing the semantics of the features contradicts the specificity of the DCA as it is important to know the source (feature) of each signal category based on its meaning. Therefore, I focused on studying the DCA data pre-processing phase while proposing new pre-processing modules. To achieve this, I used two granular computation methods known as "Rough Set Theory" and "Fuzzy Rough Set Theory". Rough set theory is described in Appendix 9.
- *Sources of the DCA sensitivity to the input class data order:* As previously stated, it was noticed that the DCA is sensitive to the input class data order. To handle this DCA issue, I proposed solutions based on the use of fuzzy set theory, a maintenance database technique, and fuzzy clustering.

Contributions, publications, and a MSc project co-supervision

- *Robust pre-processing:* In this part, I proposed new pre-processing modules based on "Rough Set Theory" and "Fuzzy Rough Set Theory". This is to build new DCA data pre-processing phases based on feature selection techniques, rather than the use of a feature reduction technique which contradicts the fundamentals of the DCA. Particularly, I have proposed different rough DCAs where each algorithm is based on a different feature selection process based on a specific immunological design [CE13d, CE12a, CE12b]. However, despite the advantages of the developed rough DCAs, as classification performance has been increased while guaranteeing the semantics, these methods have to perform data discretization. To handle this limitation, I aimed to extend the rough DCAs by proposing various fuzzy-rough DCAs where each of these algorithms is based on specific fuzzy-rough concepts for feature selection. These algorithms could minimize the information loss during data pre-processing which is due to the data quantization process [CE14d, CE14a, CE14b, CE13b, CE13e]. In summary, when focusing on the standard DCA data pre-processing phase I could develop a new data pre-processing module that replaces both the manual one, based on expert knowledge, and the module which is based on dimensionality reduction techniques.
- *Sources of the DCA sensitivity to the input class data order:* In this part, I proposed solutions based on fuzzy set theory, a maintenance database technique, and fuzzy

clustering to solve the algorithm limitation which is tied to its sensitivity to the input class data order. Specifically, the first investigations I conducted showed that the mandatory respect of the antigens class order is related to two main causes. The first cause is linked to the DCA environment which is characterized by a crisp separation between the DC immunological contexts. However, the reality is connected to imprecision by nature. Such imperfection may affect the DCA classification performance leading the algorithm being sensitive to the input class data order. Therefore, and in order to have a more robust version of [CE10] (developed in my Master Thesis) that automatically deals with the parameters settings, I have combined the fuzzy DCA version with fuzzy clustering [CE11]. The second cause tied to the DCA sensitivity is linked to the quality of the algorithm generated signal database. To deal with the quality of the signal database, the application of a maintenance method was required [CSE12]. The results of these hybridizations led to different fuzzy DCA versions which are stable classifiers [CE14b, CE14c, CE13c, CE13a]. In summary, the work conducted in this part demonstrated the possibility of combining various theories with the DCA.

- *The overall model:* Once the DCA main limitations were solved, I tended to develop a more general algorithm based on the conducted observations and conclusions [CE16b]. I have built a more robust DCA within an imprecise framework [CE16a]. The developed algorithm is based on robust and well-studied algorithmic steps. It is based on the fuzzy-rough concepts to ensure a more convenient data pre-processing phase. Indeed, it uses fuzzy set theory to ensure the stability of the algorithm. The proposed new DCA version also relies on the use of fuzzy clustering to ensure parameters' settings [CE15]. Adding to these, the algorithm applies a maintenance technique to ensure a good quality of its input parameters [DE18].

Results showed that the integrated DCA become more applicable and more adapted to real-world problems. *During my PhD defense, my published works consisted of* ***1 journal paper, and 13 peer-reviewed conference papers*** *(all of which were orally presented).* Other publications tied to my PhD work were published later (outlined in Chapter 7).

In the context of a co-supervision of a Master student, another DCA extension was developed. This new version aims at improving the DCA classification algorithmic phase which relies on a predefined set of signal values. Requiring the expert knowledge to define these values presents a limitation in many application domains. To overcome this restriction, together with my Master student Kaouther Ben Ali, we have proposed a new version of the DCA which is based on the K-Nearest Neighbors (KNN) [ACE15]. Details tied to this supervision are given in Chapter 3, Section 3.2.8.

1.4 Habilitation Thesis

In this section, I will outline the various challenges that prompted my research presented in this Habilitation Thesis. Additionally, I will provide an overview of my research approach.

1.4.1 Research challenges

Since completing my PhD Thesis in 2014, I have been engaged in a wide range of research in various domains. Some of my research builds upon the same fields I explored during my PhD, including artificial immune systems, logic and reasoning theories, but with a fresh perspective in light of new developments and challenges in the field. Other areas of my research are in new domains, such as big data, distributed systems, evolutionary computation, privacy preservation, as well as a strong focus on applying my research to real-world applications through close collaboration with partners outside of academia. This Habilitation Thesis is structured around the following three main research directions:

- **Research Direction 1: New approaches for knowledge discovery and privacy preservation using granular computation and federated learning.** In this first research direction, the use of granular computation has facilitated the development of innovative methods for big data pre-processing, and shed lights on privacy preservation using both granular computation and federated learning. These approaches, which will be explained in details in Chapter 4, cover the following research studies:
  - Big data pre-processing using rough set theory.
  - Privacy preservation for semantically enriched mobility data using granular computation and federated learning.
- **Research Direction 2: New approaches for knowledge discovery using biological computation and deep learning.** In this second research direction, different aspects from artificial immune systems, evolutionary computation, and deep learning were considered. Specifically, new approaches have been developed to adapt techniques to the big data context, while some other techniques have been adapted to various domains. These techniques will be described in Chapter 5. Example of contexts in which the proposed novel approaches lay are the following:
  - Artificial immune systems and evolutionary algorithms in scalable designs and some theoretical work.
  - A new evolutionary approach for wireless sensor network deployment.
  - New evolutionary and granular approaches for Android malware detection.
  - A new evolutionary approach for the BYOD context.

- A new multi-objective multi-agent deep interactive reinforcement learning approach

- **Research Direction 3: Personalizing healthcare: Focus on real-world applications.** In this third research direction, a special focus is given to different health real-world applications. The proposed approaches will be explained in Chapter 6, and mainly cover the following case studies:
  - A case study in epidemiology.
  - Personalized approaches to treating sepsis.

1.4.2 Research methodology

My adopted research methodology is presented in Figure 1.5. As previously noted, my research journey started in 2007 - 2008 during my Bachelor's research project. The project focused on handling uncertainty in quality control by utilizing fuzzy set theory from the field of logic and reasoning.

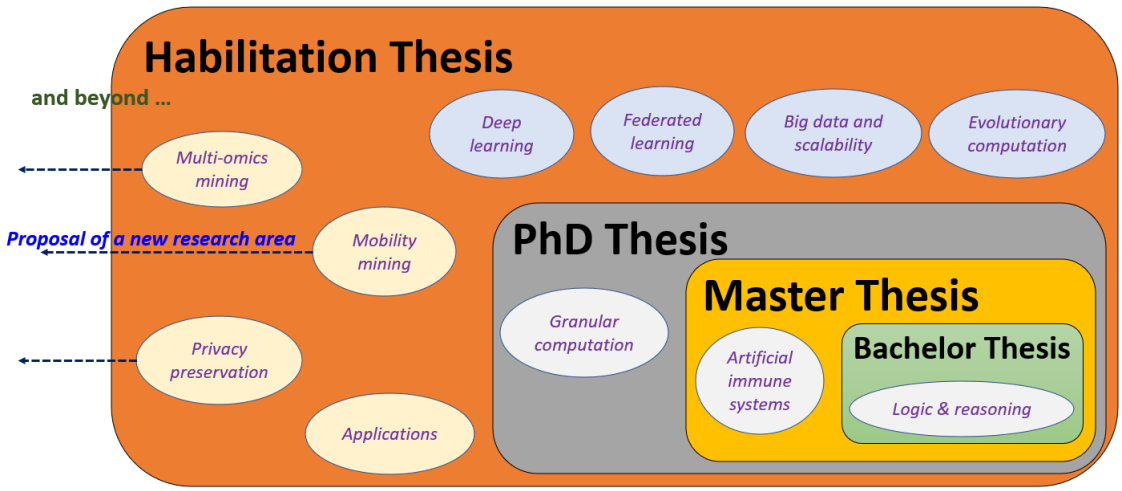

Figure 1.5: Research methodology

To work on my Master, I have extended my research topics to tackle problems from artificial immune systems using approaches based on logic and reasoning which were acquired during my Bachelor studies. For my PhD Thesis, I built novel hybrid approaches by combining the various research fields I explored during my Bachelor's and Master's studies, and broadening their scope to include granular computation. In my post-doctoral research, I tackled new challenges in areas such as scalability and distributed processing, evolutionary computation, and a strong focus on real-world applications. Additionally, I am currently focusing on mobility data mining, on multi-omics analysis and mining, and on privacy preservation. These are currently being investigated using federated learning, deep learning, and the previous acquired expertise in granular computation (among others). These domains added

new dimensions to my research fields, and enabled me to propose a new research area that I called: *"granular multiple aspect trajectory representation and privacy preservation"*.

1.5 Conclusion

Throughout this chapter, I gave an overview of my professional path and academic journey. After that, I highlighted the three main research directions that will be presented in this Habilitation Thesis. The contributions made in each of the research directions will be described in Chapter 4, Chapter 5, and Chapter 6. The next chapter will cover my aptitude for conducting research and technology transfer.

Chapter 2

In the realm of research and the development of innovation and technology transfer

2.1 Introduction

This chapter highlights my involvement and contributions in the field of research and the advancement of innovation and technology transfer. This chapter is structured as follows: Section 2.2 presents an overview of awards and honours I have obtained. Section 2.3 provides a detailed description of my involvement and contributions to the organization of scientific events and some other research activities. In Section 2.4, I present my scientific responsibilities. Section 2.5 highlights my responsibilities in the development of innovation and technology transfer. Section 2.6 lists my institutional responsibilities. Section 2.7 presents my collective responsibilities. Section 2.8 gives an overview of my publications. Finally, the chapter concludes with Section 2.9.

2.2 Honours and awards

I have obtained the following **awards**:

- Marie Sklodowska Curie Individual Fellowship (2015), MSCA Certificate (2019)
- Best Reviewer Award (iCDEc 2018)
- ACM-Women in Computing grant (2014)
- IEEE Young Researcher, First Price (EHB 2013)

Additionally, I have been **honoured to be**:

- Invited as an evaluator for the Information and Communications Technology (ICT) proposals (H2020 projects) (Big Data proposals).
- Act as a Marie Sklodowska Curie ambassador.
- Named as a female scientist role model.

- Invited as a panellist at the Parlamentarium for the debate: "Education, Research & Innovation: developing concrete synergies", Brussels, Belgium.
- Selected to be part of the Marie Skłodowska-Curie Actions-Falling Walls Lab.
- Selected to be part of the Marie Skłodowska-Curie Researchers' Night Event.
- Selected to be among the Heidelberg-Laureate-Forum most qualified young researchers to meet the Abel-Prize, Fields-Medal, the ACM-Turing award, the Nevanlinna Prize and the ACM Prize in Computing winners.
- Selected to organize a workshop with Prof Stephen Smale who was awarded the Fields Medal.

2.3 Organization of scientific events, expertise, and other research activities

2.3.1 Expertise and support to national or international organizations

- **Advisory Board Member for the European Researchers' Night (Tunisia 2018).**
- **Invited speaker** to promote project submissions to the prestigious European program Marie Skłodowska-Curie Individual Fellowship. I shared my experience with researchers during two summer schools organized by the Welsh government; one in Cardiff and the other in Swansea (2017, 2018). I presented my MSCA project as a concrete example of success and highlighted the lessons learned throughout the writing process, while providing advice to help researchers achieve their own goals.
- **Marie Skłodowska-Curie Ambassador**. My main responsibilities include (i) presenting the prestigious European MSCA program at universities and research institutes, (ii) encouraging and supporting young researchers in the development and submission of MSCA projects, (iii) sharing my experiences in the MSCA program, and (iv) providing pre-evaluation support to researchers for their MSCA research projects.
- In 2018, I was appointed as a **female scientist role model**[1] by the Marie Skłodowska-Curie Association to represent a successful model for the mobility of women scientists. Since 2018, my tasks included (i) raising awareness among women scientists about professional mobility opportunities, (ii) providing them with information on available funding programs, (iii) advising them on best practices for applying for scientific positions, (iv) encouraging their participation in international scientific projects, and (v) guiding them in building or joining international scientific networks.

[1] http://www.therolemodels.net/wp-content/uploads/2018/06/Ebook-2018-5.pdf

2.3.2 Expertise and support for public policies

- ❒ **Invited panelist** (2018), invited by the Directorate General for Education, Youth, Sport and Culture – European Commission, to contribute to the debate: "Education, Research & Innovation: developing concrete synergies"[2] in Brussels.
- ❒ **Invited speaker**, invited by the Directorate General for Education, Youth, Sport and Culture – European Commission, to contribute to the event "Marie Skłodowska-Curie Actions Stakeholders' Conference: Talents for the future: impacting careers, organizations, and systems" (03 - 04/12/2019). Together with stakeholders, we explored ways to maximize the impact of MSCA on the professional development of researchers as well as Europe's innovation potential. This discussion informed the implementation of the new Horizon Europe MSCA program.
- ❒ **Invited evaluator** (2019), by the European Commission, to assess H2020-ICT projects. I did not serve as there was a conflict of interest with submissions from Inria researchers.

2.3.3 Editorial activities

Guest Editor

- ❒ Engineering Applications of Artificial Intelligence, Special Issue on Recent Advances in Immune Computation (2018)
- ❒ Applied Soft Computing – Special Issue on Recent Advances in Immune Computation (2018)

Associate Editor

- ❒ Applied Computing and Intelligence

Regular Program Committee Member and Reviewer The Genetic and Evolutionary Computation Conference – The In'l Conf. on Advanced Intelligent Systems and Informatics – Int'l Congress on Systems Immunology, Immuno-informatics & Immune-Computation – The Special Session on Parallel and Distributed Data Mining at The International Conference on High Performance Computing & Simulation – The Int'l Conf. on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology – The ACS/IEEE Int'l Conf. on Computer Systems and Applications – Int'l Symposium on Computer Vision and the Internet – Annual Int'l Conf. on digital economy – International Symposium Advanced Artificial Intelligence in Applications – International Symposium on Rough Sets: Theory and Applications.

Reviewer in Journals Natural Computing, Springer — IEEE Transactions on Evolution-

[2] https://acmweurope.acm.org/debate-education-research-innovation-developing-concrete-synergies/

ary Computation -– IEEE Transactions on Industrial Informatics — IEEE Transactions on Cybernetics -– Journal of Systems and Software, Elsevier — Int'l Journal of Fuzzy Systems, Springer — IEEE Transactions on Emerging Topics in Computational Intelligence — Engineering Applications of artificial Intelligence, Elsevier — Swarm and Evolutionary Computation, Elsevier -– Artificial Intelligence, Elsevier — Applied Soft Computing, Elsevier -– International Journal of Approximate Reasoning, Elsevier -– Fundamenta Informaticae – Information Sciences – Artificial Intelligence Review, Elsevier – Journal of Ambient Intelligence and Humanized Computing, Springer.

Reviewer in International Conferences IEEE Symposium Series on Computational Intelligence (SSCI) – IEEE Congress on Evolutionary Computation – The Int'l Conf. on Biomedical Engineering and Sciences – The Int'l Conf. on Modelling, Simulation and Applied Optimization – The Int'l Computing Conf. in Arabic.

2.3.4 Knowledge dissemination, responsibilities, and activities within learned societies or associations

Invited talks I delivered the following as a guest speaker:

- 10/09/2023: "Federated Learning: Hype or Hope?", 21th International Conference of Complex Acute Illness, 2023 – scientific session: "Quantitative understanding of critical illness: Integrating Mechanism and Machine Learning"
- 4 - 5/05/2023: "Unleashing the Power of AI: Revolutionizing Industries and Personalizing Healthcare", Tunisian Women and DATA + AI Summit 2023, Tunis, Tunisia. The event is organized by the RECONNECTT association in France in partnership with Women in Data Science at Stanford University (WiDS).
- 08/09/2022: "Federated learning for healthcare", 20th International Conference of Complex Acute Illness
- 10/2022: "Effects of COVID-19 pandemic on education and society", Inspire Health Summit 2022
- 21/07/2022: "Seeking the fortification of the convergence between biological computation and computational biology", Dynamics of Immune Repertoires: Exploration and Translation, workshop
- 17 - 18/05/2018: 'A success story', All Wales Summer School 2018, Cardiff, Wales, UK
- 05 - 06/07/2017: 'A success story', All Wales Summer School 2017, Swansea, Wales, UK
- 06/02/2017: 'My MSC fellowship: The story so far', MSCA Campaign - From the Association to Participation, Tunis, Tunisia.

- 09 - 11/09/2015: 'Data pre-processing based on Rough Sets and the link to other theories', The International Afro-European Conference for Industrial Advancement, Villejuif, France.

Board member of EGC and co-responsable for e-EGC The Association Internationale Francophone d'Extraction et de Gestion des Connaissances (EGC), EGC school (e-EGC).

Board member of the Tunisian Artificial Intelligence Society

Evaluator for ANRT: Evaluator of a CIFRE (Industrial Agreements for Training through Research) research project submitted to ANRT (National Association for Research and Technology) (2023).

2.3.5 Organization of events and chair

- Workshop on Federated Learning (10/01/2023, Paris-Saclay, France) *(co-organizer and co-chair)*
- FedCSIS/RSTA 2022, FedCSIS/RSTA 2023 – International Symposium on Rough Sets: Theory and Applications *(co-organizer and co-chair)*
- IEEE CEC 2021, IEEE CEC 2020, IEEE CEC 2019, IEEE CEC 2018 - Special Session on Artificial Immune Systems - IEEE World Congress on Computational Intelligence *(co-organizer)*
- The first international Metaheuristics Summer School – MESS 2018, Taormina, Italy *(Publicity chair)*
- Workshop on Algorithms in Nature *(Organizer, Mentored by Prof Stephen Small (Fields Medal awardee))*, Heidelberg Laureate Forum, Germany (2017)
- 'Neural Networks and Fuzzy Logic' session at 'The Int'l Conf. on Swarm Intelligence', Harbin, China *(chair)* (2013)

2.3.6 Other activities

Research visits

- Granada Excellence Network of Innovation Laboratory, Spain (02-15/02, 2015 – 09-17/10, 2018)
- Computer Science Laboratory (LIPN), University Paris-North - 13, France (14/04 – 12/05, 2017, 25/10 – 03/11, 2017)

Secondments

- Municipality of Thira, Santorini, Greece (2022, 2023)

❒ Hôpital Hôtel Dieu, Epidemiology Department, Paris, France (2017, 2018)

❒ SEMOS Company, Quality Audit Department, Nabeul, Tunisia (2008)

❒ ESSID Company, Quality Audit Department, Nabeul, Tunisia (2008)

❒ Tunisie Télécom Company, Telecommunication Sector, Nabeul, Tunisia. (2007)

Scientific memberships

- IEEE Computer Society
- Association for Computing Machinery
- Part of the Joint European Mentoring Initiative program (JEMI)
- MCAA Academy program (mentor)
- Associate member of Euroscience
- Marie Curie Alumni Association
- AlumNode Community
- HLF-GSO-Alumni Network
- IEEE Computational Intelligence Society Task Force on Artificial Immune Systems
- IEEE Computer Society Technical Committee on Intelligent Informatics

2.4 Scientific responsibilities

2.4.1 Animation of research teams

During my tenure as an R&D Technical Project Manager at the Inria MULTISPEECH research team, I was responsible for the H2020 ICT COMPRISE project (with a budget of €3,201,016.25, see Table 2.7), where Inria acted as the coordinating institution.

- **Partners:**
  - 1 research institute in France (Inria Nancy & Inria Lille)
  - 1 university in Germany (Saarland University)
  - 4 small and medium-sized enterprises (SMEs): Netfective Technology (France), Ascora (Germany), Tilde (Latvia), and Rooter (Spain)
- **My main responsibilities involved:**
  - Leading and coordinating the consortium of over 30 researchers, research engineers, and software engineers.
  - Monitoring and tracking the scientific and technological progress of the project.

- Directing internal and external communication tasks (in collaboration with the communication office of all partners including Inria).
- Directing communications and interaction with the European Commission (in collaboration with the financial office of all partners including Inria).
- Monitoring Inria software development tasks in collaboration with the Experimentation and Development office.

2.4.2 Funded research projects

In this section, several research projects in which I have been involved and contributed to will be presented. A comprehensive list of these is given in Table 2.1. The main idea behind each research project will be described together with my key contributions in the following sections.

Table 2.1: Funded research projects

Acronym	Duration	Budget (€)	Type	Role
Prometheus	Accepted 05/2023 – duration 10 years (kickoff: first trimester of 2024)	~40 millions	Instituts Hospitalo-Universitaires (IHU)	Work package co-leader
i-RECORDS	2022 - 2025	1 430 820	ERA PerMed (European Research Area Network on Personalised Medicine)	Work package co-leader
RECORDS	2020 - 2025	9 919 695	RHU (Research Hospital University)	Work package co-leader
SEPSIS	2020 - 2025	65 000	FHU (Fédération Hospitalo-Universitaire)	Work package co-leader
MASTER	2018 - 2023	576 000	H2020-MSCA-RISE-2017	Researcher

COMPRISE	2018 - 2021	3 201 016,25	H2020-ICT-2018-20	R&D project manager
RoSTBiD-Framework	2017 - 2019	183 454,80	H2020-MSCA-IF-2015	MSCA fellow
BIG-SKY-EARTH	2015 - 2019	602 959	COST Action	Researcher
MUSES	2012 - 2015	4 673 614	FP7-ICT-2011.1.4	Researcher
EPIDEMIUM	2015 - 2016	*(unknown)*	—	Researcher

Prometheus

Description The acronym *Prometheus*[3] refers to "PRecisiOn MedicinE for healTHcare associatEd and commUnity acquired Sepsis". Its description is given in Table 2.2.

Table 2.2: Prometheus

Project duration:	10 years - kickoff: first trimester of 2024
Budget:	~40 millions €
Coordinator:	Paris-Saclay University (alongside the University of Versailles Saint-Quentin-en-Yvelines (UVSQ), the French Alternative Energies and Atomic Energy Commission (CEA), and Assistance Publique - Hôpitaux de Paris (APHP))
Type of project:	IHU
Description:	This future global center integrating research, training, and care aims to halve the mortality and sequelae caused by sepsis within ten years. Sepsis is the most severe complication of infections. It is characterized by a loss of control over inflammation, the physiological process by which the body eliminates pathogens, with this loss of control leading to vital functions impairment.

[3] https://www.universite-paris-saclay.fr/actualites/luniversite-paris-saclay-porteuse-et-co-porteuse-de-deux-instituts-hospitalo-universitaire

Despite decades of extensive research, the mechanisms underlying the dysregulation of the host's response to pathogens remain poorly understood, and no treatment has been found. The IHU PROMETHEUS has three scientific ambitions. Firstly, to better understand the interactions between the host and pathogens that lead to the progression of an uncomplicated infection to sepsis. This will help identify signatures that characterize the trajectory of each individual with an infection (endotype) and their response to a specific treatment (treatable trait). Establishing a prospective cohort will be a major support in characterizing these individual sepsis profiles. Secondly, to develop, validate, and commercialize a rapid testing platform (capable of conducting over 200 tests for endotypes and treatable traits within two hours). This will lead to the creation of a digital twin of organs and systems to assist in prompt therapeutic decision-making. Rapid decision-making, within six hours of symptom onset, is critical for treatment success. Thirdly, to develop personalized medicine.

I am the **co-leader of the work package entitled *"Mathematical modelling, Intelligent health data analytics - the path toward the digital twin"***.

i-RECORDS

Description The acronym *i-RECORDS*[4] refers to "International - Rapid rEcognition of CORticosteroiDs sensitivity or resistance in Sepsis". The description of the project is given in Table 2.3.

Table 2.3: i-RECORDS project

Project duration:	2022 - 2025
Budget:	1 430 820 €
Coordinator:	Assistance Publique - Hôpitaux de Paris (APHP)
Type of project:	ERA PerMed

[4] https://anr.fr/ProjetIA-18-RHUS-0004

Description: Sepsis and COVID-19 are both placing a major burden on societies and populations worldwide. Deregulated host response to infection is the hallmark supporting the routine use of corticosteroids (CS), a low-cost and highly efficient class of immuno-modulators, in sepsis/COVID-19. Stratifying patients based on individual immune response may improve the balance of benefit to risk of CS treatment. The project integrates different approaches to define the CS sensitivity/resistance of individual patients. The partners will elaborate signatures from different characterizations of biological systems in patients with sepsis/COVID-19. Targeted approaches will define whether characteristics at the level of DNA, RNA, proteins such as cytokines and hormones, or metabolite compounds, support predicting individual patient's CS responsiveness. Methods of artificial intelligence will integrate the high dimensional multi-level data from previous studies of this consortium and from data to be newly generated. An exploratory adaptive trial will include patients with sepsis/COVID-19 in multiple arms based on novel CS responsiveness signatures to be tested. Within each biomarker-defined cohort, patients will be randomized to receive corticosteroids or placebo allowing the evaluation of the efficiency of signatures elaborated by the partners. Signatures of CS responsiveness will be integrated for predictive enrichment of CS sensitivity and resistance of each individual patient defining personalized treatment rules, and thereby improving their chance to survive in good health. We will test the robustness of the personalized corticotherapy across subsets of patients based on gender, social categories and ethnicity.

Keywords: Sepsis, COVID-19, Corticosteroids, cytokines, hormones, machine learning

Contributions We harmonised the RECORDS and APROCCHS cohorts by establishing a standardised mapping using day-14 variables as a basis. Further discussions with the PI will determine if mapping should extend to day-28 or day-90 data. Next, we will unify the remaining cohorts (SISPCT, CORTICUS, HYPRESS, and DEXA-ARDS). The encountered obstacle is the discrepancies in variable names between the cohorts, which has increased the time required for mapping. We have also refined our AI-based signature and conducted a comprehensive statistical analysis for both the RECORDS and APROCCHS cohorts. Our subsequent step will involve conducting a unified data analysis for all cohorts. Additionally, we have worked on the development of the data management plan. For i-RECORDS, I am the **co-leader of the *"Data analysis and machine learning"* work package.**

Additionally, **I take charge of leading the project's *"data management plan"* and, in collaboration with the project coordinator, oversee its coordination.**

RECORDS

Description The acronym *RECORDS*[5] refers to "Rapid rEcognition of CS sensitive or resistant Sepsis". The description of the project is given in Table 2.4.

Table 2.4: RECORDS project

Project duration:	2020 - 2025
Budget:	9 919 695 €
Coordinator:	APHP
Type of project:	RHU
Description:	Our multidisciplinary consortium will develop point-of-care test to customize corticotherapy for sepsis at patient level within the next five years. We have built a unique sepsis cohort - APROCCHS (n=1240) - that has shown survival benefits from hydrocortisone + fludrocortisone therapy. This cohort will allow the investigation of qualitative interactions between clinical phenotypes and survival benefits or harms from CS, i.e., will allow defining sensitivity and resistance to CS. Then, using previously biobanked samples (arounnd 1000 patients) from the APROCCHS cohort, we will characterize the predictivity of GILZ expression/levels, endocan levels, SRS and polymorphism in GC-GR regulatory genes. Our consortium has already developed an assay for endocan in plasma, and will validate endocan assays from exhaled air and for its urinary metabolite. We will launch a platform trial within CRICS-TRIGERSEP (F-CRIN labeled) network, allowing to prospectively validate in an adaptive design trial, qualitative interaction between clinical response to CS and endocan, GILZ, SRS extended from transcriptomic to proteomic and metabolomic signatures, and CS induced monocytes subtypes in blood. Finally, our biomarker platform in connection with our industrial partners will bring to market point of care assays for all prospectively validated signatures, from blood, urine or exhaled air.

[5] https://anr.fr/ProjetIA-18-RHUS-0004

Keywords:	Sepsis, CS, endocan, GILZ, omics, point-of-care test

Contributions In this ongoing project, and in close collaboration with the APHP medical experts, we have developed and released an initial CS responsiveness signature. The later will be further updated and refined as new data and variables become available. This work presents the first contribution of the PhD student that I am currently co-supervising and it has been submitted to the *Journal of Computer Methods And Programs In Biomedicine*. Before recruiting the PhD student, the PostDoc and the software engineer who I have co-supervised worked on the data preparation process for the APPROCCHS and the RECORDS cohorts. This involved gaining a deep understanding of the data and aligning the variables' across different project cohorts. Additionally, together with the PhD student, we have investigated the application of a game-theoretic rough sets three-way based approach for clustering sepsis data with missing values. This contribution, was submitted to the *18th Conference on Computer Science and Intelligence Systems FedCSIS 2023*. Details about the conducted research can be found in Chapter 6 — Section 6.3. For the RECORDS project, I hold the position of the **co-leader of the *"Health Data Analytics"* work package.**

SEPSIS

Description The acronym *SEPSIS*[6] refers to "Saclay and Paris Seine Nord Endeavour to PersonnaliSe Interventions for Sepsis". The description of the project is given in Table 2.5.

Table 2.5: SEPSIS project

Project duration:	2020 - 2025
Budget:	65 000 €
Coordinator:	APHP
Type of project:	FHU
Description:	SEPSIS will fill a major gap in the management of patients with sepsis by providing point-of-care tools to individualize the diagnosis and treatment of sepsis and thereby reducing its mortality and sequels. Improving sepsis' outcomes requires early identification of sepsis and causative pathogens, and rapid characterization of the host deregulated response to infections.

[6] https://www.fhu-sepsis.uvsq.fr/

	SEPSIS will demonstrate that 1) sepsis can rapidly be recognized by cellular morphology analysis of circulating monocytes, 2) innovative diagnostic technologies for rapid pathogens identification with subsequent antibiotic susceptibility testing (ID/AST), together with prospective PK/PD monitoring of antibiotic plasma concentrations, will significantly shorten time from "sample to answer" for pathogen ID/AST, and result in decreased treatment failures, 3) combined artificial intelligence and multi-omics guided interventions will markedly increase the benefit to risk ratio of hemodynamic, renal, metabolic and immune management. Taken together, these three steps will provide the basis for customized interventions for sepsis. Then, we will translate this characterization into innovative rapid bedside tests that will predict clinical outcomes and guide therapies in patients with sepsis.
Keywords:	Infection, organs failure, translational research, big data, trials

Contributions At present, our first contribution is the initial CS responsiveness signature which has been released as part of the RECORDS project and integrated into the APHP information system. As new data and variables become available, this signature will be updated and refined. Although the initial stages of the SEPSIS and RECORDS projects are the same, the data and investigations will diverge in subsequent stages. For the SEPSIS project, I am the **co-leader of the *"Artificial intelligence for decision making"* work package**.

MASTER

Description The acronym *MASTER*[7] refers to "Multiple ASpects TrajEctoRy management and analysis". The description of the project is given in Table 2.6.

Table 2.6: MASTER project

Project duration:	2018 - 2023
Budget:	576 000 €
Coordinator:	National Research Council (Consiglio Nazionale delle Ricerche), Italy
Type of project:	H2020-MSCA-RISE-2017

[7] http://www.master-project-h2020.eu/consortium/

Description:	An ever-increasing number of diverse, real-life applications, ranging from mobile phone calls to social media and land, sea, and air surveillance systems, produce massive amounts of spatio-temporal data representing trajectories of moving objects. Trajectories, commonly represented by sequences of timestamps and position coordinates, thanks to the high availability of contextual and semantic-rich data can be enriched and are evolving to more comprehensive and semantically significant objects. In the MASTER project, we envision holistic trajectories, meaning trajectories characterized by the fact that the spatio-temporal and semantic aspects are intimately correlated and should be considered as a whole. However current state of art does not provide management and analysis methods "ready for use" for these multiple aspects trajectories. The overarching objective of this project is to form an international and inter-sectoral network of partners working on a joint research programme by developing methods to build, manage and analyse multiple aspects trajectories. These methods are driven by application scenarios from three different domains: tourism, sea monitoring and public transportation.
Keywords:	Mobility data analysis, trajectory data, social media, machine learning, big data, tourism, transportation, sea monitoring

Contributions In this project, I have established a new research area that I called *"granular multiple aspect trajectory representation and privacy preservation"*. A position paper about this area was published at the *4th International Symposium on Rough Sets: Theory and Applications (RSTA'22)*, part of the 17*th Conference on Computer Science and Intelligence Systems (FedCSIS)* [DB22]. Details about the conducted research can be found in Chapter 4 — Section 4.3.2. MASTER is a Research and Innovation Staff Exchange (RISE) project, and I hold the position of **a researcher**.

COMPRISE

Description The acronym *COMPRISE*[8] refers to "Cost-effective, Multilingual, Privacy-driven voice-enabled Services". The description of the project is given in Table 2.7.

Table 2.7: COMPRISE project

[8] https://cordis.europa.eu/project/rcn/218720/factsheet/en

Project duration:	2018 - 2021
Budget:	3 201 016,25 €
Coordinator:	National Institute for Research in Digital Science and Technology (Inria)
Type of project:	H2020-ICT-2018-20. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) — Research and Innovation action — ICT-29-2018 - A multilingual Next Generation Internet
Description:	Besides visual and tactile, the Next Generation Internet will rely more and more on voice interaction. This technology requires huge amounts of speech and language data in every language to reach state-of-the-art performance. The standard today is to store the voices of end users in the cloud and label them manually. This approach raises critical privacy concerns, it limits the number of deployed languages, and it has led to market and data concentration in the hands of big non-European companies such as Google, Facebook, etc. COMPRISE defines a fully private-by-design methodology and tools that will reduce the cost and increase the inclusiveness of voice interaction technology. This leads to a holistic easy-to-use software development kit. The sustainability of this new ecosystem will be demonstrated for three sectors with high commercial impact: smart consumer apps, e-commerce, and e-health.
Keywords:	Spoken interaction, privacy-by-design, deep learning.

Contributions My main responsibilities are described in Section 2.4.1.

RoSTBiDFramework

Description The acronym *RoSTBiDFramework*[9][10] refers to "Optimised Framework based on Rough Set Theory for Big Data Pre-processing in Certain and Imprecise Contexts". The description of the project is given in Table 2.8.

[9] http://rostbid.dcs.aber.ac.uk/

[10] https://cordis.europa.eu/project/rcn/207793/factsheet/en

Table 2.8: RoSTBiDFramework project

Project duration:	2017 - 2019
Budget:	183 454,80 €
Coordinator:	Aberystwyth University, UK
Type of project:	H2020-EU.1.3.2. - Nurturing excellence by means of cross-border and cross-sector mobility — MSCA-IF-2015-EF - Marie Skłodowska-Curie Individual Fellowships (IF-EF) — MSCA-IF-EF-ST - Standard EF
Description:	Over the last decades, it has become difficult to quickly acquire the most useful information from the huge amount of data at hand. Thus, it is necessary to perform data (pre-)processing as a first step. In spite of the existence of many techniques for this task, most of the state-of-the-art methods require additional information for thresholding and are neither able to deal with the big data veracity aspect nor with their computational requirements. This project's overarching aim is to fill these major research gaps by developing big data pre-processing approaches based on Rough Set Theory (RST).
Keywords:	Big data, rough set theory, missing values

Contributions Chapter 4 (Section 4.2) and Chapter 6 (Section 6.2) present the main contributions made in the context of my MSCA project. Based on the knowledge acquired during this project, I have also made other contributions in other research areas, mainly in developing new artificial immune systems in big data. This is highlighted in Chapter 5 (Section 5.2.2).

In total, this project resulted in 3 publications in the *IEEE Big Data Conference (IEEE Big Data)* [DZBL17][DZB+18][HDSZ18], in 1 publication at the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)* [DZS+18], in 1 publication at the *Genetic and Evolutionary Computation Conference (GECCO)* [Dag18a], in 1 publication at the *La Conférence Extraction et Gestion des Connaissances (EGC)* [DZBL18b], and in 1 publication at the EGC workshop *Fouille de Donneée Complexes (FDC)* [DZBL18a]. It also resulted in 4 journal publications: in the *Knowledge and Information Systems Journal* [DZBL20], in *Fundamenta Informaticae* [CDZ21], in the *Artificial Intelligence Review Journal* [CDAB21], and in the *Swarm and Evolutionary Computation Journal* [Dag18b].

BIG-SKY-EARTH

Description The acronym *BIG-SKY-EARTH*[11] refers to "Big Data Era in Sky and Earth Observation". As presented in Table 2.9, BIG-SKY-EARTH, is a European project which has received funding from the European Cooperation in Science and Technology (COST). The project involves partners from 28 countries.

Table 2.9: BIG-SKY-EARTH project

Project duration:	2015 - 2019
Budget:	602 959 € (covering four other COST actions)
Chair:	Dr. Dejan Vinkovic, Science and Society Synergy Institute, Croatia
Type of project:	COST Action
Description:	The challenges related to data volume, variety and velocity are similar in astronomy and Earth observations, with computer science as the common denominator. The BIG SKY EARTH Action aims at boosting the communication within and between disciplines by identifying and clustering relevant common solutions developed within research and industrial environments.
Keywords:	Astronomy, earth observations, big data, visualization, visual analytics, astroinformatics, geoinformatics

Contributions As an **involved researcher** in this project, and in close collaboration with a researcher from the *Department of Computer Science and Engineering, Saints Cyril and and Methodius University, North Macedonia*, we have worked on a book chapter entitled "When Evolutionary Computing meets Astro and Geo- Informatics". The chapter is part of the *Knowledge Discovery in Big Data from Astronomy and Earth Observation*[12] book [DM20]. Details about this work can be found in Chapter 5 — Section 5.2.2.

MUSES

Description The acronym *MUSES*[13] refers to "Multiplatform Usable Endpoint Security". The description of the project is given in Table 2.10.

[11] https://www.cost.eu/actions/TD1403/#tabs|Name:overview

[12] https://rb.gy/7a2elt

[13] https://cordis.europa.eu/project/rcn/105550/factsheet/en

Table 2.10: MUSES project

Project duration:	2012 - 2015
Budget:	4 673 614 €
Coordinator:	S2 Grupo de Innovación en Procesos Organizativos SL, Spain
Type of project:	FP7-ICT - Specific Programme "Cooperation": Information and communication technologies — ICT-2011.1.4 - Trustworthy ICT — CP - Collaborative project (generic)
Description:	Data security and privacy are of fundamental importance to organizations, where they are defined and managed via security policies. Most security incidents are caused by organization insiders, either by their lack of knowledge or inadequate or malicious behaviour. Nowadays information is highly distributed amongst corporate servers, the cloud and multiple personal devices like PDAs, tablets and smart phones. These are not only information holders but also user interfaces to access corporate information. Besides, the Bring Your Own Device practice is becoming more common in large organisations, posing new security threats and blurring the limits between corporate and personal use. In this situation enforcement of security policies is increasingly difficult, as any strategy with a chance to succeed must take into account several changing factors: information delocalisation, access from heterogeneous devices and mixing of personal and professional activities. The overall purpose of MUSES is to foster corporate security by reducing the risks introduced by user behaviour.
Keywords:	Corporate security, Bring Your Own Device, evolutionary algorithms, machine learning

Contributions Chapter 5 — Section 5.4 presents one of the contributions made in MUSES. As an **involved researcher**, the project allowed me to collaborate with leading researchers and a PhD student.

EPIDEMIUM

Description The aim of EPIDEMIUM is to address cancer research related scientific questions. It is about gathering people from various backgrounds — from data science to medicine, from ethics to design — and mastering complementary skills to tackle cancer research. The rise of the project (also considered as a programme) in 2015 is based on a

joined effort of both “Roche”[14] and “La Paillasse”[15].

Contributions With respect to the aim of EPIDEMIUM, as an **involved researcher**, and in close collaboration with several researchers and epidemiologists at Hôpital Hôtel-Dieu de Paris (hospital), we performed a case study where the application of the Sp-RST method discussed in Chapter 4 — Section 4.2.2 was investigated. This is to simplify the learned model for epidemiologists when it comes to colon cancer incidence prediction. The results of this case study were published in the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2018 (ECML-PKDD)* [DZS+18]. Further details about this case study can be found in Chapter 6 — Section 6.2.

2.5 Responsibilities in the development of innovation and technology transfer

As a “Partnership and Innovation Project Lead” within the “Technology Transfer, Innovation and Partnership department” at Inria (2019 - 2020), I contributed to the following achievements:

- ❒ I facilitated the connection between Inria researchers and industrial partners (e.g., connecting NOMEXY with the ORPAILLEUR team). In doing so, I helped establish industrial partnerships while supporting and assisting Inria researchers with technology transfer issues and challenges.
- ❒ I participated in the negotiation and execution of contracts of various types (e.g., European Grant Agreement, industrial CIFRE).
- ❒ In order to protect the research results of Inria researchers (intellectual property, licenses, etc.), I participated in the creation of a license.
- ❒ I assisted Inria researchers in drafting patents (initial phase).
- ❒ I participated in supporting Inria researchers in the creation of startups (e.g., Cybi) (initial phase).
- ❒ I assisted Inria researchers in the development of European projects.
- ❒ I participated in conferences/congress/events to promote Inria’s technological offerings to businesses (e.g., Digital Excellence Forum @ICT Proposer Days 2019, 360 Possibles, Impression Deeptech, Salon Cité Santé).
- ❒ In order to provide the best support to Inria researchers in writing ERC projects, I participated in several information sessions.

[14] https://www.roche.fr/

[15] https://lapaillasse.org/

- More recently, as co-leader for a work package in the RECORDS project (see Table 2.4), we developed a corticosteroid response prediction model and successfully deployed it as a web service within the APHP's information system.

2.6 Institutional responsibilities

I am engaged in various institutional responsibilities. These responsibilities include pedagogical tasks, which are detailed in Chapter 1, Sections 1.2.1 and 1.2.2, as well as active participation in doctoral mentorship training and research support programs, which are discussed in Chapter 3, Section 3.4. Additionally, I have been actively involved in doctoral support and awareness-raising activities, which are also covered in Chapter 3, Section 3.5. My supervision activities are also covered in Chapter 3. In addition to these, I have taken on other responsibilities to complement my contributions, which include:

- Member of the Selection Committee - "Maître de conférences" position (GALAXIE) — University Evry, Paris-Saclay. (2023)
- Served as a member of the PhD admission commission – sciences and technologies of information and communication (STIC) Doctorial school, Paris-Saclay. (2023)
- Jury member of the "Doctoral students prize from the Saclay plateau" Paris Saclay, France. (2022, 2023)
- UVSQ (2020 – current): Jury member of MSc deliberation and Jury member of the BSc examination and deliberation committee (president, assessor).
- ISG (2013 - 2017): Jury member of the BSc examination and deliberation committee and Jury member of MSc and BSc vivas.
- FSEGN (2010 - 2013): Jury member of the BSc examination and deliberation committee and Jury member of BSc vivas.
- Development of report templates (in LaTeX) for writing BCs project reports that are considered as standards at FSEGN[16].
- Development of report templates (in LaTeX) for writing master's theses and doctoral dissertations that are considered as standards at ISG Tunis.

2.7 Collective responsibilities

I am involved in various administrative and collective responsibilities. I actively participate in the activities and seminars of the Paris-Saclay graduate school "Sciences and technologies of information and communication", the DAVID laboratory, as well as the seminars of the

[16] http://www.fsegn.rnu.tn/?page=accueil&section=archives&step=1&range=10

ADAM research team, and contribute to the overall life of the department. For instance, I am the Co-responsible for the website renewal of the DAVID laboratory. I am collaborating with two members of DAVID on the redesign of the website. We are responsible for drafting the specifications and overseeing the Junior Enterprise team in charge of the technical implementation of the project.

2.8 Overview of publications

My published works encompass a variety of topics, including knowledge discovery in databases, approximated reasoning, artificial immune systems, evolutionary computation, machine learning, big data, scalability, and KDD applications for knowledge extraction. Overall, my published works consist of **14 journal papers, 1 book, 3 book chapters, 31 peer-reviewed conference and workshop papers, and 9 seminar proceedings and press articles.** To provide a comprehensive overview of my publications, Figure 2.1 presents a categorized summary of my publications per year and per category, excluding submissions and 2023 publications. Figure 2.2 highlights the rankings of the publications[17] based on SJR[18] and CORE[19].

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
Journal papers						1	2			2	3	2	4	14
Book											1			1
Book chapters							3							3
Conference papers	1	1	3	7	4	3		1	6		1	2	2	31
Seminar proceedings and press articles					1			4	4					9
Total	1	1	3	7	5	4	5	5	10	2	5	4	6	58

Figure 2.1: Summary of publications

	Ranking				
	Q1	Q2	Q3	*Unranked*	Total
Journal papers	7	1	3	3	14
	Q2	Q4			
Book chapters	1	2			3
	A	B	C		
Conference papers	7	9	6	9	31

Figure 2.2: Description of publications' rankings

[17] It is to be noted that the International Conference of Artificial Immune Systems (ICARIS), in which I published in 2010, 2011, and 2012, joined the Genetic and Evolutionary Computation Conference (GECCO) – ranked **A** –, since 2013.

[18] https://www.scimagojr.com/

[19] https://www.core.edu.au/

2.9 Conclusion

This chapter provided a comprehensive overview of my research and technology transfer activities and achievements. In the next chapter, my aptitude for supervision will be presented.

Chapter 3

Aptitude for supervision

3.1 Introduction

This chapter primarily demonstrates my experience in supervision, encompassing my involvement in supporting students' academic and research pursuits.This chapter is structured as follows: in Section 3.2, a description of my different supervisions will be presented. In Section 3.3, an overall view of the list of publications with co-supervised and collaborated students will be highlighted. Section 3.4 presents my participation in doctoral mentorship training and research support programs. Section 3.5 elucidates my doctoral support and awareness-raising activities. Finally, the conclusion is given in Section 3.6.

3.2 Supervisions

Since 2010, I have been actively engaged in the guidance and mentorship of students at several institutions. My experience includes supervising and co-supervising students pursuing Bachelor's, Master's, and Doctoral degrees, as well as a postdoctoral researcher and a software engineer at FSEGN and ISG, Tunisia, and at UVSQ, Paris-Saclay. Table 3.1 presents an overview of my supervision accomplishments at these institutions. Table 3.2 shows the

current positions of some of the students I have co-supervised.

Table 3.1: Summary of supervisions

Total number of students	**Level**	**Period**	**University**	**Detailed number of students**
4	**PhD**	2022 - current	UVSQ, Paris-Saclay	2
		2022 - current	Sorbonne Paris North University*	1
		2016 – 2020	ISG ***(defended on 12/2021)***	1
1	**PostDoc**	2022	UVSQ, Paris-Saclay	1
1	**Software Engineer**	2020 - 2021	UVSQ, Paris-Saclay	1
3	**MSc**	2020 - 2021	FSEGN	1
		2018 – 2019	ISG	1
		2014 – 2015	ISG	1
22	**M2 Datascale****	2022 - 2023	UVSQ, Paris-Saclay	17
		2021 - 2022	UVSQ, Paris-Saclay	5
27	**BCs**	2013 - 2015	ISG	9
		2010 – 2013	FSEGN	18

* *I am currently co-supervising Kodjo Mawuena M Amekoe together with Dr Hanene Azzag and Prof Mustapha Lebbah; yet, my role is not declared as an official and formal co-supervisor.*

** *Students taking the "Data Management in Large-Scale Distributed Systems" master module.*

Table 3.2: Current positions of some students I have co-supervised

Level	**Current position**

PhD ***(defended on 12/2021)*** *(details in Section 3.2.4)*	Permanent high school teacher at Dar Chaabane El Fehri High-school, Nabeul, Tunisia, and a temporary university teacher at IT Business School, and at FSEGN, Tunisia
PostDoc *(details in Section 3.2.5)*	Associate Professor at college of engineering and technology, American university of the middle east, Kuwait
Software Engineer *(details in Section 3.2.5)*	Data Engineer, CGI, France
MSc *(details in Section 3.2.7)*	Functional consultant Microsoft Dynamics 365 CE & Power Platform, Javista, France

In what follows, I will give an overview of the different research projects I have co-supervised.

3.2.1 PhD Thesis 1

Institute:	UVSQ, Paris-Saclay
Laboratory:	DAVID, ADAM Group
Doctoral School	Sciences and technologies of information and communication (STIC)
Enrollment:	2022 - current
Thesis title:	Machine learning based approaches for multi-omics data in personalized treatment of sepsis
Funding schema	RHU RECORDS project (PIA project)
Co-supervision:	75%
Co-supervision with:	Karine Zeitouni (25%)
Student:	Rahma Hellali

Thesis topic: This thesis is part of the RHU RECORDS project (Chapter 2 – Table 2.4) which aims to identify and validate biomarkers predicting the therapeutic response to corticosteroids in the context of sepsis. The project is based on three types of omics, – genomics, transcriptomics, and metabolomics – which are among the most robust and by far the most used in the exploration of human diseases. The search for biomarkers (not only omics) predicting the response to corticosteroids in sepsis presents a special focus in this Thesis. Specifically, the main objective of this thesis is to study the existing machine learning techniques in the analysis of omics data obtained within the framework of the RHU RECORDS, and in discovering novel biomarkers, and to propose new methods according to the limitations that will be identified.

Keywords: Machine learning, Dimensionality reduction, Multi-omics data, Personalized treatment of sepsis.

Publications:

1. **Rahma Hellali**, *Zaineb Chelly Dagdia*, **Ahmed Ktaish**, Karine Zeitouni, and Djillali Annane. "Corticosteroid sensitivity detection in sepsis patients using a personalized data mining approach: a clinical investigation." Computer Methods And Programs In Biomedicine. [***Journal paper (submitted)***]
2. **Rahma Hellali**, *Zaineb Chelly Dagdia*, and Karine Zeitouni. "Application of a Game-Theoretic Rough Sets Three-way based Approach for Clustering Corticosteroids Responsiveness in Sepsis Patients." 18th Conference on Computer Science and Intelligence Systems FedCSIS 2023. [***Conference paper (submitted)***]

3.2.2 PhD Thesis 2

Institute: UVSQ, Paris-Saclay

Laboratory: DAVID, ADAM Group

Doctoral School STIC

Enrollment: 2022 - current

Thesis title:	FL4Mobility A federated learning approach for privacy of semantically enriched mobility data
Funding schema	LabEx Digicosme & École supérieure d'informatique, électronique, automatique (ESIEA)
Co-supervision:	25%
Co-supervision with:	Karine Zeitouni (25%)
	Nazim Agoulmine (25%)
	Bassem Haidar (25%)
Student:	Saloua Bouabba
Thesis topic:	The main aim of this PhD proposal is to investigate, study, and apply Federated Learning in the context of privacy preserving mobility mining; with a special focus on semantically enriched trajectories. This will consider the main analysis tasks, such as trajectory clustering and mobility prediction.
Keywords:	Federated Learning, Mobility mining, Semantic trajectory.

3.2.3 PhD Thesis 3

Institute:	Sorbonne Paris North University
Laboratory:	Laboratoire d'Informatique de Paris-Nord (LIPN)
Doctoral School	Galilée doctoral school
Enrollment:	2022 - current
Thesis title:	Adaptive Learning applied to Fraud
Funding schema	CIFRE (Banque Populaire Caisse d'Epargne - BPCE)
Co-supervision:	I am currently co-supervising Kodjo Mawuena M Amekoe together with Dr Hanene Azzag and Prof Mustapha Lebbah; yet, my role is not declared as an official and formal co-supervisor.
Co-supervision with:	Hanene Azzag (50%)
	Mustapha Lebbah (50%)
Student:	Kodjo Mawuena M Amekoe

Thesis topic:	The fraudulent nature of a transaction can be classified as a binary outcome, which banks determine by analyzing customer feedback from credit card transactions or communications from other financial institutions related to check transactions. This research project addresses two interconnected aspects. Firstly, it focuses on the challenge of imbalanced data, a common issue in fraud detection. Secondly, it places great importance on the practical application of the developed models in detecting fraudulent activities, particularly in terms of explainability.
Keywords:	Fraud detection, Imbalanced data, Deep learning.

3.2.4 PhD Thesis 4

Institute:	High Institute of Management of Tunis, Tunisia
Laboratory:	Strategies for Modelling and ARtificial inTelligence Laboratory (SMART-LAB)
Enrollment:	2016 - 2020 ***(defended on 12/2021)***
Thesis title:	Android Malware Detection based on Real and Artificial Patterns using Evolutionary and Rough Set Approaches
Co-supervision:	75%
Co-supervision with:	Prof. Slim Bechikh (25%)
Student:	Manel Jerbi
Thesis topic:	Mobile devices have become an essential part of daily life, allowing for easy access to different applications and services. However, the rapid growth of mobile internet technology has led to increased cyber threats, including malware. Anti-virus scanners are unable to fully protect against this diverse range of threats, and the high availability of attacking tools and anti-detection techniques has made malware attacks accessible to anyone, regardless of skill level.

The need for accurate malware detection methods is crucial for both individuals and businesses as even one attack can result in significant data compromise and losses. Obfuscation techniques used by malware developers hinder manual and automated inspections and make it difficult for detection tools to accurately identify malware. This thesis proposes new and effective malware detection methods through the use of innovative techniques from evolutionary algorithms and rough set theory.

Keywords: Android malware detection, API call sequences, Artificial malicious patterns, Detection rules generation, Evolutionary algorithm, Bi-level optimization, Rough set theory.

Publications:

1. **Jerbi Manel**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Android malware detection as a bi-level problem." Computers & Security 121 (2022): 102825. ***[Journal paper]***
2. **Jerbi Manel**, *Zaineb Chelly Dagdia*, Slim Bechikh, & Lamjed Ben Said, "On the Use of Artificial Malicious Patterns for Android Malware Detection". Journal of Computers & Security: 92: 101743 (2020). ***[Journal paper]***

* **Jerbi Manel**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "A Novel Malware Detection Approach based on Three-way Decisions and Bi-level Optimization." Cognitive Computation ***[Journal paper (submitted)]***

3. **Manel Jerbi**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Malware Evolution and Detection Based on the Variable Precision Rough Set Model." In 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 253-262. IEEE, 2022. ***[Conference paper]***
4. **Manel Jerbi**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said, "Malware Detection Using Rough Set Based Evolutionary Optimization", Proceedings of the 28th International Conference on Neural Information Processing, ICONIP'2021, Bali, Indonisia, Springer, pp. 634-641. ***[Conference paper]***

Current position	Permanent high school teacher at Dar Chaabane El Fehri High-school, Nabeul, Tunisia, and a temporary university teacher at IT Business School, and at FSEGN, Tunisia.

3.2.5 PostDoc and Software Engineer

Institute:	UVSQ, Paris-Saclay
Laboratory:	DAVID, ADAM Group
Contract:	2022 (PostDoc)
	2020 - 2021 (Software Engineer)
Funding schema	RHU RECORDS project (PIA project)
Co-supervision:	50%
Co-supervision with:	Karine Zeitouni (50%)
Postgraduates:	Hassan Moustafa Harb (PostDoc)
	Ahmad Ktaish (Software Engineer)
Description of work:	The PostDoc and the software engineer worked on the data preparation process of the APPROCCHS and RECORDS cohorts. This involved gaining a deep understanding of the data, aligning the variables, utilizing machine learning to create preliminary corticosteroid signatures, and deploying the selected signatures within the Assistance Publique–Hôpitaux de Paris (APHP) information system.
Publication:	1. **Rahma Hellali**, *Zaineb Chelly Dagdia*, **Ahmed Ktaish**, Karine Zeitouni, and Djillali Annane. "Corticosteroid sensitivity detection in sepsis patients using a personalized data mining approach: a clinical investigation." Computer Methods And Programs In Biomedicine. [***Journal paper (submitted)***]
Current positions	(Hassan Moustafa Harb) Associate Professor at college of engineering and technology, American university of the middle east, Kuwait

(Ahmed Ktaish) Data Engineer, CGI, France

3.2.6 MSc. Thesis 1

Institute:	High Institute of Management of Tunis, Tunisia
Laboratory:	Strategies for Modelling and ARtificial inTelligence Laboratory (SMART-LAB)
Enrollment:	2020 - 2021
Master title:	Evolutionary Induction of Regression Trees for Predictive Analytics
Co-supervision:	75%
Co-supervision with:	Prof. Slim Bechikh (25%)
Student:	Khaled Sethom
Master topic:	The primary objective of the MSc thesis project is to develop an evolutionary based approach for the induction of regression trees. This approach will leverage the principles of evolutionary algorithms to guide the search for optimal tree structures.
Keywords:	Regression, Evolutionary computation, Optimization.

3.2.7 MSc. Thesis 2

Institute:	High Institute of Management of Tunis, Tunisia
Laboratory:	SMART-LAB
Enrollment:	2018 - 2019
Master title:	Many-Objective Optimization of Wireless Sensor Network Deployment
Co-supervision:	75%
Co-supervision with:	Prof. Slim Bechikh (25%)
Student:	Omar Ben Amor

Master topic:	An efficient deployment of Wireless Sensor Network (WSN) has become a leading area of research. Practical scenarios related to WSN deployment are often considered as optimization models with multiple conflicting objectives simultaneously enhanced. Previously, it had been shown that moving from mono- to multi-objective resolution of WSN deployment is beneficial. However, since the deployment of real-world WSNs encompasses more than three objectives, a multi-objective optimization may harm other deployment criteria which are conflicting with the selected ones. Thus, our aim is to go further and explore the modeling and the resolution of WSN deployment in a many-objective fashion.
Keywords:	Many-objective optimization, Wireless sensor network, Deployment model, Evolutionary algorithms.
Publications:	1. **Ben Amor Omar**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Many-objective optimization of wireless sensor network deployment." Evolutionary Intelligence (2022): 1-17. ***[Journal paper]***
Current position	Functional consultant Microsoft Dynamics 365 CE & Power Platform, Javista, France

3.2.8 MSc. Thesis 3

Institute:	High Institute of Management of Tunis, Tunisia
Laboratory:	LAboratory of Operation Researches, DEcision and process Control (LARODEC)
Enrollment:	2014 - 2015
Master title:	A New Version of the Dendritic Cell Immune Algorithm based on the K-Nearest Neighbors
Co-supervision:	90%
Co-supervision with:	Prof. Zied Elouedi (10%)
Student:	Kaouther Ben Ali

Master topic: Many researches have demonstrated the promising DCA classification results in many real-world applications. Despite of that, it was noticed that some information should be given by the expert in order to perform its classification task. To classify a new data item, the expert knowledge is required to calculate a set of signal values. Indeed, to achieve this, the expert has to provide some specific formula capable of generating these values. Yet, the expert mandatory presence has received criticism from researchers. Therefore, in order to overcome this restriction, we have proposed a new version of the DCA combined with the K-Nearest Neighbors (KNN). KNN is used to provide a new way to calculate the signal values independently from the expert knowledge.

Keywords: Artificial immune systems, K-Nearest Neighbors, Classification.

Publications:

1. **Kaouther Ben Ali**, *Zeineb Chelly*, and Zied Elouedi, "A New Version of the Dendritic Cell Immune Algo rithm based on the K Nearest Neighbors". Proceedings of the 22th International Conference on Neural Information Processing, ICONIP'2015, Istanbul, Turkey, Springer, pp 688 695. ***[Conference paper]***

3.2.9 M2 DataScale students

Institute: UVSQ, Paris-Saclay

DataScale Module: Exploratory and Predictive Data Mining

Number of M2 students: 17 students (2022 - 2023)

5 students (2021 - 2022)

Projects' titles: Investigating feature selection techniques to improve data mining tasks (2022 - 2023)

Investigating a three-way clustering approach for handling missing data (2021 - 2022)

Supervision: 100%

Description of work: The aim of these research projects is to provide students with a comprehensive introduction to the field of research. This involves exposing students to new and emerging areas of research and helping them to develop a deep understanding of the underlying principles of the associated research papers. Through this experience, students will be able to understand the technical aspects of the algorithms and their implementation, as well as replicate the results presented in the research papers. The ultimate goal is to help students build critical thinking skills by identifying any limitations in the algorithms, and to inspire them to propose innovative solutions to improve the algorithms. These projects are designed to be hands-on and interactive, providing students with a rich and engaging experience that will help them to develop the skills and knowledge required for a successful career in research.

3.3 Summary of publications with co-supervised and collaborated students

The list of publications with my students who I have co-supervised or collaborated with includes a total of 5 peer-reviewed journal articles, and 10 conference papers.

➽ **Journal papers**

✍ ***Publications resulting from co-supervision***

1. **Ben Amor Omar**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Many-objective optimization of wireless sensor network deployment." Evolutionary Intelligence (2022): 1-17. ***[MSc. student]***
2. **Jerbi Manel**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Android malware detection as a bi-level problem." Computers & Security 121 (2022): 102825. ***[PhD student]***
3. **Jerbi Manel**, *Zaineb Chelly Dagdia*, Slim Bechikh, & Lamjed Ben Said, "On the Use of Artificial Malicious Patterns for Android Malware Detection". Journal of Computers & Security: 92: 101743 (2020). ***[PhD student]***

➦ ***Submissions***

① **Rahma Hellali**, *Zaineb Chelly Dagdia*, **Ahmed Ktaish**, Karine Zeitouni, and Djillali Annane. "Corticosteroid sensitivity detection in

sepsis patients using a personalized data mining approach: a clinical investigation." Computer Methods And Programs In Biomedicine. ***[PhD student, Software Engineer]***

② **Jerbi Manel**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "A Novel Malware Detection Approach based on Three-way Decisions and Bi-level Optimization." Cognitive Computation ***[PhD student]***

✍ ***Publications resulting from collaborative work with PhD students***

4. *Zaineb Chelly Dagdia*, Md Shamsuzzoha Bayzid, and **Pavel Avdeyev**, "Biological computation and computational biology: survey, challenges, and discussion". Artificial Intelligence Review, 54(6): 4169-4235 (2021). ***[PhD student]***

5. *Zaineb Chelly Dagdia*, Christine Zarges, **Gael Beck**, and Mustapha Lebbah, "A Scalable and Effective Rough Set Theory based Approach for Big Data Pre-processing" Knowledge and Information Systems: 62: 3321-3386 (2020). ***[PhD student]***

➽ **Conference papers**

✍ ***Publications resulting from co-supervision***

1. **Manel Jerbi**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Immune-Based System to Enhance Malware Detection". IEEE 2023 Congress on Evolutionary Computation. **(to appear)** ***[Postdoctoral research]***

2. **Manel Jerbi**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said. "Malware Evolution and Detection Based on the Variable Precision Rough Set Model." In 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 253-262. IEEE, 2022. ***[PhD student]***

3. **Manel Jerbi**, *Zaineb Chelly Dagdia*, Slim Bechikh, and Lamjed Ben Said, "Malware Detection Using Rough Set Based Evolutionary Optimization", Proceedings of the 28th International Conference on Neural Information Processing, ICONIP'2021, Bali, Indonisia, Springer, pp. 634-641. ***[PhD student]***

4. **Kaouther Ben Ali**, *Zeineb Chelly*, and Zied Elouedi, "A New Version of the Dendritic Cell Immune Algo rithm based on the K Nearest Neighbors". Proceedings of the 22th International Conference on Neural Information Processing, ICONIP'2015, Istanbul, Turkey, Springer, pp 688 695. ***[MSc. student]***

➦ ***Submission***

③ **Rahma Hellali**, *Zaineb Chelly Dagdia*, and Karine Zeitouni. "Application of a Game-Theoretic Rough Sets Three-way based Approach for Clustering Corticosteroids Responsiveness in Sepsis Patients." 18th Conference on Computer Science and Intelligence Systems FedCSIS 2023. ***[PhD student]***

✍ ***Publications resulting from collaborative work with PhD students***

5. **Paloma de las Cuevas**, Pablo Garcia-Sanchez, *Zaineb Chelly Dagdia*, Maria Isabel Garcia-Arenas, and Juan Julian Merelo, "Automatic rule extraction using Genetic Programming in a Bring Your Own Device scenario". The Leading European Event on Bio Inspired Computation, EVOSTAR'2020, Seville, Spain, pp. 54-69. ***[PhD student]***

6. *Zaineb Chelly Dagdia*, Christine Zarges, **Gael Beck**, Hanene Azzag and Mustapha Lebbah, "A Distributed Rough Set Theory Algorithm based on Locality Sensitive Hashing for an Efficient Big Data Pre-processing". Proceedings of the IEEE Big Data Conference, Proceedings of the IEEE Big Data Conference, BigData'2018, Seattle, USA, IEEE, pp. 2597-2606. ***[PhD student]***

7. **Azam Hamidinekoo**, *Zaineb Chelly Dagdia*, **Zobia Suhail**, and Reyer Zwiggelaar, "Distributed Rough Set Based Feature Selection Approach to Analyse Deep and Hand-crafted Features for Mammography Mass Classification", Proceedings of the IEEE Big Data Conference, BigData'2018, Seattle, USA, IEEE, pp. 2423-2432. ***[PhD student]***

8. *Zaineb Chelly Dagdia*, Christine Zarges, **Gael Beck**, and Mustapha Lebbah, "Modèle de Sélection de Caractéristiques pour les Données Massives", Proceedings of the 15ème édition de l'atelier Fouille de Données Complexes, FDC'2018, Paris, France, pp 1-12. ***[PhD student]***

9. *Zaineb Chelly Dagdia*, Christine Zarges, **Gael Beck**, and Mustapha Lebbah, "Nouveau Modèle de Sélection de Caractéristiques basé sur la Théorie des Ensembles Approximatifs pour les Données Massives", Proceedings of the 18ème édition de la conférence internationale francophone Extraction et Gestion de Connaissances, EGC'2018, Paris, France, pp 377-378. ***[PhD student]***

10. *Zaineb Chelly Dagdia*, Christine Zarges, **Gael Beck**, and Mustapha Lebbah, "A Distributed Rough Set Theory based Algorithm for an Efficient Big Data Pre-processing under the Spark Framework". Proceedings of the IEEE Big Data Conference, BigData'2017, Boston, USA, IEEE, pp 911-916. ***[PhD student]***

3.4 Participation in doctoral mentorship training and research support programs

As part of my professional development in doctoral supervision, I participated in the **"Doctoral Mentorship Training Program"** *(Formation à l'encadrement des doctorant(e)s)* organized by Paris-Saclay on the 13th, 14th, and 23rd of September 2022. The program was 21 hours in duration and provided me with valuable insights and hands-on learning experiences to strengthen my supervision abilities. Through engaging lectures and interactive discussions, I gained a deeper understanding of best practices in supervision and mentorship. Additionally, I had the opportunity to network with other experienced supervisors.

In addition to my experience in the doctoral context, I have also sought out opportunities to expand my skills in other areas of research support. One such opportunity was the **"How to Support Researchers in Writing H2020 Proposals"** workshop, which was organized by Lorraine University and animated by Dr Sean McCarty[1]. The workshop took place on March 12th, 2020, and provided me with valuable insights and training in the art of supporting researchers in writing successful proposals for the Horizon 2020 program. Through this training, I was able to gain a deeper understanding of what makes a proposal successful, and what factors are taken into consideration by grant review committees.

3.5 Doctoral support and awareness-raising activities

I am actively participating in a range of activities aimed at enhancing the academic experience of doctoral students, postgraduates, and raising awareness among the wider public. In what follows, I will provide a more detailed overview of some of these activities.

As a dedicated supporter of master and doctoral students, **I actively participate in the Doctoral School activities** by giving workshops (e.g., efficient use of LaTeX for the production of high-quality technical and scientific documentations). In addition, I have being **a jury member for the "Doctoral students prize from the Saclay plateau"**; (2022, 2023) further emphasizing my commitment to the advancement and recognition of doctoral students in their research endeavors.

For postgraduates, and as a **MSCurie ambassador,** I have been actively involved in **several organizational and outreach initiatives aimed at promoting the Horizon 2020 and now the Horizon MSC Individual Fellowships among *postgraduates*.** Through my talks and presentations, I aim to encourage postgraduates to consider applying for postdoctoral positions and to inspire them to apply for the MSC Individual Fellowships.

I was honored to be invited by the European Commission to **contribute to the development of the new Horizon Europe program for the MSC Individual Fellowships**.

[1] http://www.hyperion.ie/seanmccarthy.htm

This was via participating in the “Marie SkŁdowska-Curie Actions Stakeholders’ Conference: Talents for the Future: Impacting Careers, Organizations, and Systems,” held on December 3rd and 4th, 2019. My involvement in this conference further highlights my commitment to promoting postdoctoral opportunities and the MSC Individual Fellowships.

I have also been a **speaker at two Summer Schools** organized by the Welsh Government, sharing my success story with the MSC Fellowship, one held in Cardiff in 2018 and the other in Swansea in 2017, UK. In these talks, I emphasized the benefits of pursuing a postdoctoral position, including opportunities for professional growth, gaining expertise in a specific field, and preparing for a future academic or industry career. Additionally, I was invited as a **speaker at the European Commission’s MSCA Campaign “From the Association to Participation"** in Tunis, Tunisia.

Since the completion of my MSC fellowship in 2019, I have been actively engaged in **mentorship and support of fellow researchers**. Specifically, I have being serving as a **mentor for the MCAA Academy program within the MSCA Alumni community.** The MCAA Academy is a mentorship initiative designed to help early-career researchers develop their skills and advance their careers. By providing guidance and support to junior researchers, I aim to pay forward the benefits I received from my MSC Fellowship and help the next generation of researchers reach their full potential.

My dedication to inspiring and encouraging young researchers extends beyond postgraduates and doctoral students to all levels of education. This is demonstrated by my participation and **contributions to various outreach initiatives**, such as the **Marie SkŁodowska-Curie Actions Falling Walls Lab** and the **Marie SkŁodowska-Curie Researchers’ Night** events in 2018. At these events, I had the opportunity to share my research with a wide range of audiences, including pupils at primary and secondary schools, as well as higher education students. My goal was to inspire and encourage these young individuals to pursue their interests and careers in science and research, regardless of their current educational level. Through my presentations and interactions with the audience, I aimed to showcase the excitement and impact of scientific research, as well as to provide guidance and encouragement to those who are considering a career in this field.

Since my MSC fellowship, I have embraced the role of a **female scientist role mode**l, actively promoting and advocating for the participation and advancement of women in the field of science and technology. One of my outreach activities in this regard is my **keynote speaking engagement at the “Tunisian Women and DATA + AI Summit 2023”**. The summit, which was held on May 04 & 05 2023 in Tunis, Tunisia, was organized in partnership with WiDS (Women in Data Science of Stanford University) and is focused on promoting the integration of young people and women in the digital world.

Also, as a **board member of the Tunisian Artificial Intelligence Society**, my role involves actively contributing to the strategic decision-making process and the overall direction

of the society. I collaborate with fellow board members to develop and implement initiatives that promote the advancement and adoption of artificial intelligence (AI) in Tunisia. This includes organizing conferences, workshops, and seminars to facilitate knowledge sharing and networking among AI professionals and enthusiasts. Additionally, I participate in outreach activities to raise awareness about the potential and ethical implications of AI, fostering collaborations with academia, industry, and government institutions to drive AI research and innovation in Tunisia.

In addition to my public speaking engagements and mentorship initiatives, my outreach activities also extend to **the publication of press releases to promote the participation and advancement of young researchers.** One of these press releases, entitled **"Role Models for Mobility of Women Scientists"**[2] was **published by the MSC Actions (MSCA)** to encourage young women scientists to embrace mobility and seek new professional experiences. Another press release I wrote was: **"Experience, Learn and Share at the Heidelberg Laureate Forum"**. It aimed to encourage young researchers to participate in the Heidelberg Laureate Forum. This press release was **published by both the Heidelberg Laureate Forum Media**[3] and **ACM's Women in Computing (ACM-W) Europe**[4]. Through these press releases, I have been able to reach a wider audience and inspire young researchers to embrace new challenges, opportunities, and experiences. I believe that these initiatives can play a crucial role in helping young scientists develop their skills and knowledge, establish new connections and collaborations, and advance their careers.

3.6 Conclusion

Throughout this chapter, I have tried to showcase my experience in guiding and supporting students at various academic levels, including undergraduates and postgraduates. Now to delve into the research conducted in my Habilitation Thesis, in the next chapter, the first research direction will be presented.

[2] http://www.therolemodels.net/wp-content/uploads/2018/06/Ebook-2018-5.pdf

[3] https://scilogs.spektrum.de/hlf/experience-learn-share-heidelberg-laureate-forum/

[4] https://acmweurope.acm.org/experience-learn-and-share-at-the-heidelberg-laureate-forum/

Part II

Summary of scientific contributions

Chapter 4

New approaches for knowledge discovery and privacy preservation using granular computation and federated learning

4.1 Introduction

In a wide variety of fields, data are being collected at an intense pace and processed thanks to the latest high technologies and available distributed systems. Big data which became easily accessible rise many challenges specifically when it comes to data pre-processing and privacy preservation. Focusing on these challenges is the main scope of the current first research direction.

This chapter is structured as follows: In Section 4.2, the motivation and contributions made in developing distributed granular computing based approaches for feature selection are described. In Section 4.3, initial contributions and work in progress in privacy-aware analysis and preservation are presented. Finally, a conclusion is given in Section 4.4.

4.2 Big data pre-processing using rough set theory

This section describes the conducted research in the frame of my RoSTBiDFramework H2020-MSCA-IF funded project (described in Chapter 2 – Table 2.8).

4.2.1 Context and motivation

Data reduction, and specifically feature selection, is a crucial step in the knowledge discovery in databases process. This step presents an important point of interest as many real-world applications may have a very large number of features [WZWD14]. Feature selection is a challenging process due to the very large search space that reflects the combinatorially large number of all possible feature combinations to select from. This task is becoming more difficult as the total number of attributes is increasing in many big data application domains combined with the increased complexity of those problems.

In the context of big data, it is worth mentioning that a detailed study was conducted in

[BCRFPB+18] where authors performed a deep analysis of the scalability of the state-of-the-art feature selection techniques that belong to the filter, the embedded, and the wrapper techniques. In [BCRFPB+18], it was demonstrated that the state-of-the-art feature selection techniques will obviously have scalability issues when dealing with big data. Authors have proved that the existent techniques will be inadequate to handle a high number of attributes in terms of training time and/or effectiveness in selecting the relevant set of features. Thus, the adaptation of feature selection techniques for big data problems seems essential and it may require the redesign of these algorithms and their incorporation in parallel and distributed environments/frameworks. One potential solution is the MapReduce paradigm [DG10], which provides a robust and efficient framework for addressing big data analysis. Several works were concentrated on parallelizing and distributing machine learning techniques using MapReduce [VCR+16, ZWCG17, ZGWC17]. Additionally, since 2015, several more flexible paradigms have emerged to extend the standard MapReduce approach, notably Apache Spark [SD15]. Apache Spark has demonstrated successful applications in numerous real-world data mining and machine learning problems [SD15].

With the aim of choosing the most relevant and pertinent subset of features, a variety of feature reduction techniques[1] were proposed within the Apache Spark framework to deal with big data in a distributed way. Among these are several feature extraction methods and very few feature selection techniques. To further expand this restricted research, some other feature selection techniques were proposed in literature which are based on evolutionary algorithms [PdRRG+15]. Nevertheless, most of these techniques suffer from some shortcomings. For instance, they usually require the user or expert to deal with the algorithms' parametrization, where some other techniques simply order the attributes set and let the user choose his/her own subset. There are some other feature selection techniques that require the user to indicate how many attributes should be selected. All of these are counted as significant drawbacks as they require users to make a decision based on their own (possibly subjective) perception. To overcome the shortcomings of the state-of-the-art techniques, it seemed crucial to look for a filter approach that does not require any external or supplementary information to function properly. Rough Set Theory (RST) [TP09], as an efficient granular computation theory, detailed in Appendix 9, can be used as such a technique.

The use of RST in data mining and knowledge discovery has proved to be very successful in many application domains such as in classification [Lin01], clustering [Lin02] and in supply chain [BS10]. This success is explained by the several aspects of the theory in dealing with data. For example, the theory is able to analyse the facts hidden in data, does not need any supplementary information about the given data such as thresholds or expert knowledge on a particular domain and is also capable to find a minimal knowledge representation [DG00]. This is achieved by making use of the granularity structure of the provided data only.

[1] https://spark.apache.org/docs/2.2.0/ml-features.html

Although algorithms based on rough sets have been widely used as efficient filter feature selectors, most of the classical rough set algorithms are sequential ones, computationally expensive, and can only deal with non-large data sets. The prohibitive complexity of these algorithms comes from the search for an optimal attribute sub-set through the computation of an exponential number of candidate subsets. This is quite impractical for big data sets as it becomes clearly unmanageable to build the set of all possible combinations of features.

To overcome these limitations, in my MSCA project, I have considered three contexts namely, the **certain context** where the big data veracity characteristic was discarded, i.e., no missing attribute values where considered, the **uncertain context** where two types of missing data were dealt with specifically *"do not care"* values and *"lost"* values, and the **optimized context** where an optimized partitioning of the feature search space was considered. Based on these three main contexts, different contributions reflecting new rough set based approaches for big data feature selection within a distributed framework will be presented in what follows.

4.2.2 Contributions

Context 1: The certain context

Firstly, big data in a certain context was handled. Here, the big data veracity aspect was not dealt with, i.e., the presence of all attribute values was assumed (no missing values).

Problem statement As presented in Appendix 9, the classical rough set theory for feature selection presents an exhaustive search as the theory needs to compute every possible combination of attributes. The number of the possible generated attribute sub-set of combinations with m attributes from a set of N total attributes is $\binom{N}{m} = \frac{N!}{m!(N-m)!}$ [GE03]. Thus, the total number of feature subsets to generate is $\sum_{i=1}^{N} \binom{N}{i} = 2^N - 1$. For example, for $N = 30$ we have roughly 1 billion combinations. This constraint prevents us to use high dimensional data sets as the number of feature subsets is growing exponentially in the total number of features N. Data can be too big that their size can easily exceed the available random-access memory. These are the main motivations for our proposed rough set based solution which makes use of parallelization.

Contribution To overcome the standard RST inadequacy to perform feature selection in the certain context of big data, a parallel rough set based algorithm, named "Sp-RST", was developed. A high level description of the algorithm is presented in Figure 4.1, whereas a detailed description of the approach is given in what follows:

- The Sp-RST input big database refers to the data stored in the Distributed File System (DFS). To perform distributed tasks on the given DFS, a Resilient Distributed Dataset (RDD) is built. The latter can be formalized as a given information table defined as T_{RDD}. T_{RDD} is defined via a universe $U = \{x_1, x_2, \ldots, x_N\}$ which refers to the set

of data instances (items), a conditional feature set $C = \{c_1, c_2, \ldots, c_V\}$ that includes all the features of the T_{RDD} information table and finally via a decision feature D of the given learning problem. D refers to the label (also called class) of each T_{RDD} data item and is defined as follows: $D = \{d_1, d_2, \ldots, d_W\}$. C presents the conditional attribute pool from where the most significant attributes will be selected.

- To ensure the scalability of Sp-RST when dealing with a large number of attributes, the algorithm first partitions the input T_{RDD} information table into a set of m data blocks based on splits from the conditional feature set C. This task leads to the following formalization: $T_{RDD} = \bigcup_{i=1}^{m}(C_r)T_{RDD_{(i)}}$; where $r \in \{1, \ldots, V\}$. r defines the number of attributes that will be considered to build every $T_{RDD_{(i)}}$ data block. Based on this, every $T_{RDD_{(i)}}$ is built using r random attributes which were selected from C, where $\forall T_{RDD_{(i)}} : \nexists\{c_r\} = \bigcap_{i=1}^{m} T_{RDD_{(i)}}$.
- Based on the adopted Apache Spark formalism, a reformulation of the standard RST main feature selection concepts, i.e., the Lower Approximations (LA) (i.e., feature definitely included) and Upper Approximations (UA) (i.e., feature possibly included), was defined. Specifically, new equivalence relation formula for each of LA and UA, which allow the partitioning of the universe into smaller subsets, were defined. Then, the updated LA and UA concepts were implemented with respect to the partitioning of big data into different blocks. This was achieved by defining a set of input key/value pairs that are then processed to merge these values together to form a possibly smaller set of values. This is based on the *map* and *reduce* functions.
- With respect to the Sp-RST parallel implementation design, the distributed Sp-RST algorithm will be applied to every $T_{RDD_{(i)}}$ while gathering all the intermediate results from the distinct m created partitions; rather than being applied to the complete T_{RDD} that encloses the whole set C of conditional features. Based on this design, we can ensure that the algorithm can perform its feature selection task on a computable number of attributes and therefore overcome the standard rough set computational inefficiencies.
- The outcome of each created partition can be either only one reduct $RED_{i_{(D)}}(C_r)$ or a set (a family) of reducts $RED^{F}_{i_{(D)}}(C_r)$. As previously highlighted in Appendix 9, any reduct among the $RED^{F}_{i_{(D)}}(C_r)$ reducts can be selected to describe the $T_{RDD_{(i)}}$ information table. Therefore, in case where Sp-RST generates a single reduct for a specific $T_{RDD_{(i)}}$ partition then the final output of this attribute selection phase is the set of features defined in $RED_{i_{(D)}}(C_r)$. These attributes represent the most informative features among the C_r features, and generate a new reduced $T_{RDD_{(i)}}$ defined as: $T_{RDD_{(i)}}(RED)$. The latter reduced base guarantees nearly the same data quality as its corresponding $T_{RDD_{(i)}}(C_r)$ which is based on the full attribute set C_r. In the other case where Sp-RST generates multiple reducts then the algorithm performs a

random selection of a single reduct among the generated family of reducts $RED^{F}_{i_{(D)}}(C_r)$ to describe the corresponding $T_{RDD_{(i)}}$. This random selection is supported by the RST fundamentals and is explained by the same level of importance of all the reducts defined in $RED^{F}_{i_{(D)}}(C_r)$. More precisely, any reduct included in the family of reducts $RED^{F}_{i_{(D)}}(C_r)$ can be selected to replace the $T_{RDD_{(i)}}$ (C_r) attributes.

Variants: At this step, two versions were developed when a family of reducts is generated: (i) either Sp-RST performs a random selection of a single reduct among $RED^{F}_{i_{(D)}}(C_r)$ as described above, or (ii) based on the user selected parameter, Sp-RST generates the *core* from $RED^{F}_{i_{(D)}}(C_r)$ and uses it as the final outcome.

- At this level, the output of every i data block is $RED_{i_{(D)}}(C_r)$ which refers to the selected set of features. Nevertheless, since every $T_{RDD_{(i)}}$ is described using distinct attributes and with respect to $T_{RDD} = \bigcup_{i=1}^{m}(C_r)T_{RDD_{(i)}}$, a union operator on the generated selected attributes is needed to represent the original T_{RDD}. This is defined as $Reduct_m = \bigcup_{i=1}^{m} RED_{i_{(D)}}(C_r)$.

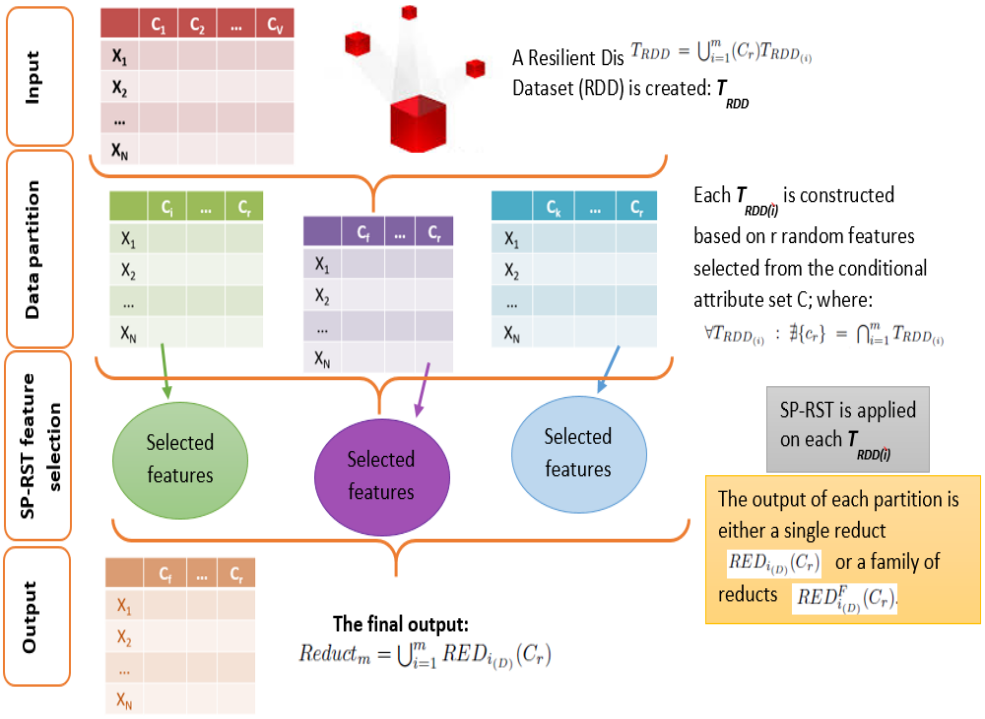

Figure 4.1: A high level description of Sp-RST

To further guarantee the Sp-RST feature selection performance while avoiding any critical information loss, to evolve the algorithm and to refine it, Sp-RST runs over N iterations on the T_{RDD} m data blocks. Through all these N iterations, Sp-RST will first randomly build the m distinct $T_{RDD_{(i)}}$ as explained above. Once this is achieved and for each partition, the algorithm's distributed tasks will be performed. This will result in N $Reduct_m$. Therefore, finally, an intersection operator applied on all the obtained $Reduct_m$ is required. This

is defined as $Reduct = \bigcap_{n=1}^{N} Reduct_m$. Sp-RST could diminish the dimensionality of the original data set from $T_{RDD}(C)$ to $T_{RDD}(Reduct)$ by removing irrelevant features at each computation level.

Key conclusions In [DZBL17, DZBL20], we have shown that using a moderate number of nodes yields a considerably improvement of the runtime and good speedup performance of Sp-RST. Results also showed that Sp-RST is competitive or better against other feature selection methods and only induces very small information loss. Moreover, Sp-RST is able to reliably identify the most and least important features in the data set, which can be an important aspect when interpreting the feature selection results from an application perspective.

Context 2: The uncertain context

While data are being collected, a number of different factors enters into play that can easily affect data quality. Among these factors are noise, environmental factors (e.g., humidity, temperature, etc.), human error in measurements, a lack of response in scientific experiments, system failures, low quality of sensors, problems of data transfer in digital systems, respondents' unwillingness to respond to survey questions, and many more. Therefore, in such circumstances, the presence of missing values in the collected databases is inevitable.

The second focus on my MSC project is handling big data in an uncertain context where missing values govern the used data. In this specific context, distributed RST based techniques dealing with two types of missing data, namely *"do not care"* and *"lost"* values, were developed. *"do not care"* attribute values can be defined as values which were not recorded because they were irrelevant, or redundant or not necessary to make a decision or to classify a case. An example of an irrelevant case can be a nurse who was able to diagnose a patient without some medical equipment. In practice, the presence of do not care attribute values, or conditional attribute values with respect to the rough set terminology, means that initially the data item was classified (the decision attribute value was assigned) in spite of the fact that the attribute value was not given. This is possible since the remaining attribute values were sufficient for such a classification or to make a decision. On the other hand, *"lost"* attribute values are values that matter but they are missing. They can be defined as values which were not recorded in the information table because they were forgotten. It is also possible that such values were originally recorded in the data set but later on they were mistakenly erased. Another possibility can be that a respondent refuses to answer a survey question, and hence a missing value — referred to as *lost* value — is present in the information table.

Problem statement As previously highlighted, data in real world are rarely clean. The presence of missing values in databases can dramatically degrade the results of interpretation of data sets. They can lead to a wrong prediction or classification and can also cause a high

bias for any given model being used. Missing values can also diminish the quality for any performance metric. Therefore, dealing with missing values is an important issue in the domain of data mining and knowledge discovery in databases. In current state-of-the-art approaches, missing values in data are typically addressed by removal, replacement, or imputation methods. However, these methods may not be suitable for all data instances, as they can potentially compromise the integrity of estimations by discarding valuable information present in data items with missing values. This can introduce bias into the estimation results. Consequently, it becomes essential to explore alternative approaches that can handle this problem by conducting feature selection on the original unaltered big data.

Contribution A novel scalable rough set based feature selection technique within an imprecise context, named "Miss-RST", was developed. To perform its feature selection task, the main LA and UA concepts used in the Sp-RST implementation [DZBL17, DZBL20] were redefined to deal with "lost" and "do not care" missing values. This was achieved with respect to the fundamentals defined in [GB08]. In [GB08], the author studied incomplete decision tables from a rough sets perspective, and defined (among other concepts) characteristic relations which are generalization of the indiscernibility relation. In my MSC work, these concepts were again adapted to function in a scalable manner. Technically, Miss-RST functions as follows:

- The Miss-RST input incomplete big database refers to the data stored in the DFS. To perform distributed tasks on the given DFS, an RDD is built. The latter can be formalized as a given incomplete information table defined as T_{RDD}. T_{RDD} is defined via a universe $U = \{x_1, x_2, \ldots, x_N\}$, a conditional feature set $C = \{c_1, c_2, \ldots, c_V\}$, and finally via a decision feature D of the given learning problem. D refers to the label of each T_{RDD} data item and is defined as follows: $D = \{d_1, d_2, \ldots, d_W\}$. All decision values are specified, i.e., they are not missing. C presents the conditional attribute pool from where the most significant attributes will be selected. The different values that an attribute c_i, where $i \in \{1, \ldots, V\}$, can have are as follows: $\{c, ?, *\}$, where "c" refers to any possible non-missing value that an attribute can take, "?" refers to *lost* values, and "$*$" refers to *do not care* values. Each case x_i, where $i \in \{1, \ldots, N\}$ is defined by at least one non-missing attribute value.

- To ensure the scalability of our proposed algorithm when dealing with a large number of attributes, Miss-RST first partitions the input T_{RDD} incomplete information table into a set of m data blocks based on splits from the conditional feature set C. This task leads to the following formalization: $T_{RDD} = \bigcup_{i=1}^{m}(C_r)T_{RDD_{(i)}}$; where $r \in \{1, \ldots, V\}$. r defines the number of attributes that will be considered to build every $T_{RDD_{(i)}}$ data block. Based on this, every $T_{RDD_{(i)}}$ is built using r random attributes which were selected from C, where $\forall T_{RDD_{(i)}} : \nexists\{c_r\} = \bigcap_{i=1}^{m} T_{RDD_{(i)}}$, and $\forall T_{RDD_{(i)}} : \exists\{c_r\} \notin \{*, ?\}$.

- A reformulation of the LA and UA main feature selection concepts were defined in

a distributed way. Specifically, the characteristic relations defined in [GB08] were redesigned. Characteristic relations are interpreted as the smallest set of cases that are indistinguishable from x_i using all attributes from C_r .

- The rest of the Miss-RST parallel implementation design follows the same logic as the one defining Sp-RST [DZBL17, DZBL20].

Key conclusions In our experiments, the initial results indicated that the scalable Miss-RST technique performs competitively when compared to other baseline models that utilize imputation-based methods. We also conducted a sensitivity analysis, varying the proportion of missing values in an artificially generated dataset. Even as the proportion of missing values increased, Miss-RST showcased its ability to maintain the quality of feature selection while minimizing information loss. These findings affirm the effectiveness of the scalable Miss-RST technique in handling internally missing values without compromising the overall efficacy of the feature selection process.

Context 3: The optimized context

The theory of rough sets goes through the calculation of the dependency of attributes, $\gamma(C, c_i)$, that it can perform feature selection. With this aim, and as a first process, the indiscernibility relation, defined as $IND(P)$, for all attribute has to be calculated. $IND(P)$ searches for similar attributes values and gathers the corresponding features to form the set of the equivalence relations. With respect to these fundamental RST notions, it is essential to guarantee data dependency in order to define the most reliable and consistent equivalence relations, so that the most representative reduct set can be guaranteed. Nevertheless, assuring data dependency is considered as a big challenge when it comes to distributed environments and parallel computing.

In this third focus of my MSC project, the partitioning of the large feature search space is optimized within the distributed environment. In this concern, new rough set based feature selection approaches have been developed within an optimized framework.

Problem statement As described in Section 4.2.2, Sp-RST [DZBL17, DZBL20] applies a random process when partitioning the feature search space; a process that does not guarantee data dependency. More specifically, Sp-RST partitions the information table T into m data blocks T_i based on splits from the conditional attribute set C in a way that: $T = \bigcup_{i=1}^{m}(C_r)T_i$; where $r \in \{1, \ldots, V\}$. r is a user defined parameter that refers to the number of attributes which will be considered to build each T_i data block and V refers to the total number of attributes. Each T_i data block is built based on r features which are randomly selected from the conditional attribute set C. Each constructed partition is processed independently in the distributed environment so that all of the intermediate results can be gathered at the end from the various m partitions. Based on this Sp-RST architecture and implementation

design, it is very probable that similar attributes will be part of different partitions T_i. Consequently, an erroneous estimation of the constructed $IND(P)$ is more likely to occur. More precisely, the applied random process may mislead the RST feature selection process by generating a non-relevant reduct. To deal with this challenge, new optimized distributed RST approaches have been developed.

Contributions In this section, two contributions will be presented. The first contribution elucidates our RST based solution that we named "LSH-dRST". The proposed solution makes use of the Locality Sensitive Hashing (LSH) algorithm [GIM99] for a more intelligent and reliable partitioning of the universe. The second contribution describes a work that is based on bi-clustering algorithms rather than hashing techniques to better guarantee the partitioning of the universe with respect to its two dimensions.

Contribution 3.1: Description In order to deal with complete big data sets and to make use of the LSH adopted technique, and within a distributed environment, the appropriate set of LSH buckets is first generated. After that, these generated buckets will be mapped into several partitions. Then, the entire rough set feature selection process will be partitioned into different elementary tasks where each of these will be executed independently on each generated bucket. As a last step, the intermediate results will be conquered to finally acquire the final output, i.e., the reduct set. A high level description of LSH-dRST is presented in Figure 4.2, whereas a detailed description of the approach is given in what follows:

- To make LSH-dRST scalable with the large number of attributes and with respect to data dependency, the given T_{RDD} information table is partitioned into B data blocks using the B generated LSH buckets. The different buckets correspond to splits from the conditional feature set C and each bucket covers a definite feature space incorporating all similar and close data instances based on their attribute values. This can be formalized as: $T_{RDD} = \bigcup_{b=1}^{B}(C_h)T_{RDD_{(b)}}$; where $h \in \{1, \ldots, V\}$. The parameter h refers to the value which is generated by LSH, and corresponds to the number of attributes per bucket that will be considered to build each $T_{RDD_{(b)}}$ data block.
- Once the buckets are defined, each $T_{RDD_{(b)}}$ is partitioned into S sub-information tables Cl based on the K nearest neighbors approach. K corresponds to the number of attributes per sub-information table and on which LSH-dRST will be applied. This can be formalized as: $T_{RDD_{(b)}} = \bigcup_{s=1}^{S} Cl_s(K)$; where $S = C_h/K$.
- Aiming at ensuring scalability, instead of applying LSH-dRST to T_{RDD} which covers the whole conditional attribute set C, the distributed LSH-dRST different tasks will be applied to every single $Cl_s(K)$, where $s \in \{1, \ldots, S\}$. At the end, all the intermediate results will be congregated from the different Cl sub-information tables of every single $T_{RDD_{(b)}}$ partition. Based on such a process, we can ensure that LSH-dRST can be applied to a computable and manageable number of attributes while preserving data

dependency.

- Each $Cl_s(K)$ will have an output which can be in one of the following two forms: (i) either a single reduct $RED_{s_{(D)}}(K)$ or (ii) a family of reducts $RED^F_{s_{(D)}}(K)$. Accordingly, in case where LSH-dRST generates a single reduct, for a specific $Cl_s(K)$ sub-information table, then the output of this attribute selection phase is the set of the $RED_{s_{(D)}}(K)$ attributes. These selected attributes reflect the most informative features among the initial K features defining $Cl_s(K)$. This results in a new reduced $Cl_s(K)$, defined as $Cl_s(RED)$, which preserves nearly the same data quality as its corresponding $Cl_s(K)$ which is based on the whole feature set K. The second case is when LSH-dRST generates a family of reducts. In this particular case, the algorithm will randomly select one reduct among $RED^F_{s_{(D)}}(K)$ to represent the corresponding $Cl_s(K)$. This means that any reduct which is included in $RED^F_{s_{(D)}}(K)$ can be used to replace the K features of $Cl_s(K)$.

  Variants: At this step, two versions have been developed when a family of reducts is generated: (i) either LSH-dRST performs a random selection of a single reduct among $RED^F_{s_{(D)}}(K)$ as described above, or (ii) based on the user selected parameter, LSH-dRST generates the *core* from $RED^F_{s_{(D)}}(K)$ and uses it as the final outcome.

- At this stage, each Cl_s sub-information table has its output $RED_{s_{(D)}}(K)$ which corresponds to the selected attributes. Nevertheless, as each Cl_s is defined using distinct features within different $T_{RDD_{(b)}}$ feature search spaces and with respect to $T_{RDD_{(b)}} = \bigcup_{s=1}^{S} Cl_s(K)$, a union operation of the generated selected features is required. This is to represent the initial T_{RDD}, defined as $Reduct = \bigcup_{b=1}^{B} \bigcup_{s=1}^{S} RED_s$.

  By removing the set of irrelevant attributes, LSH-dRST can reduce the dimensionality of the data, specifically the large number of features, from $T_{RDD}(C)$ to $T_{RDD}(Reduct)$.

Contribution 3.1: Key conclusions Our results in [DZB+18, CDZ21] revealed that the number of buckets does not have a significant influence on the speedup, but the number of sub-information tables F does. In terms of *the number of features*, it is very concentrated around its median, implying a low variance in the number of features selected. For Sp-RST the results reported in [DZBL17, DZBL20] were much more erratic with no clear tendency based on the parameter setting. Finally, we observed that LSH-dRST has at least comparable performance to all other feature selection techniques and outperforms some of them. Moreover, we noticed that the classification results are quite stable with respect to the parameter settings in LSH-dRST.

Contribution 3.2: Description LSH-dRST aims at optimizing the feature search space that it can guarantee data dependency. However, as it is based on a hashing technique, the approach only considers a single dimension of the information table to make the partitioning of the universe. Such hashing based approaches succeed when only some subset of features is

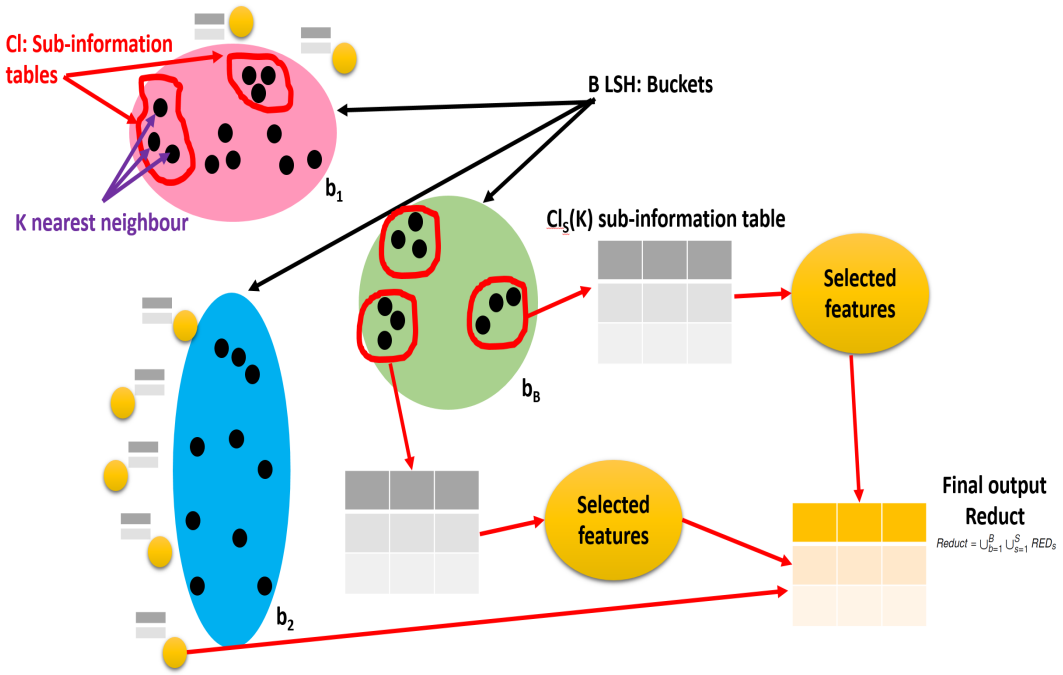

Figure 4.2: A high level description of LSH-dRST

important to a specific data block. However, it is crucial to be able to capture local patterns within a data set as both rows and columns subsets may contain elements that are not necessarily adjacent to each other.

With respect to this challenge, bi-clustering techniques come into play. By applying bi-clustering techniques, it will be possible to capture heterogeneous patterns that manifest only in subsets of features and subsets of data items. A presentation of the skeleton of the method is given in Figure 4.3 accompanied with the following description:

- To make the method scalable with the large number of attributes and with respect to data dependency in its two dimensions, the given T_{RDD} information table is partitioned into p bi-clusters (or data blocks) $B_k = (I_k, J_k)$, where $k = \{1, 2, \ldots, p\}$ using a bi-clustering technique. $I_k \subseteq U$, are the data items (rows) of the k^{th} bi-cluster, and $J_k \subseteq C$, are the features (columns) of the k^{th} bi-cluster; where each of the bi-clusters meets some homogeneity criteria [MO04]. The different p bi-clusters correspond to splits from the conditional feature set C and the universe U. This can be formalized as: $T_{RDD} = \bigcup_{k=1}^{p}(I_k, J_k)B_{(k)}$.
- Aiming at ensuring scalability, instead of applying the method to T_{RDD} which covers the whole conditional attribute set C, the distributed method different tasks will be applied to every single $B_{(k)}$ bi-cluster. At the end, all the intermediate results will be congregated from the different p bi-clusters. Based on such a process, we can ensure

that the method can be applied to a computable and manageable number of attributes while preserving data dependency in its two dimensions.

- The rest of the method's parallel implementation design follows the same logic as the one defining Sp-RST [DZBL17, DZBL20].

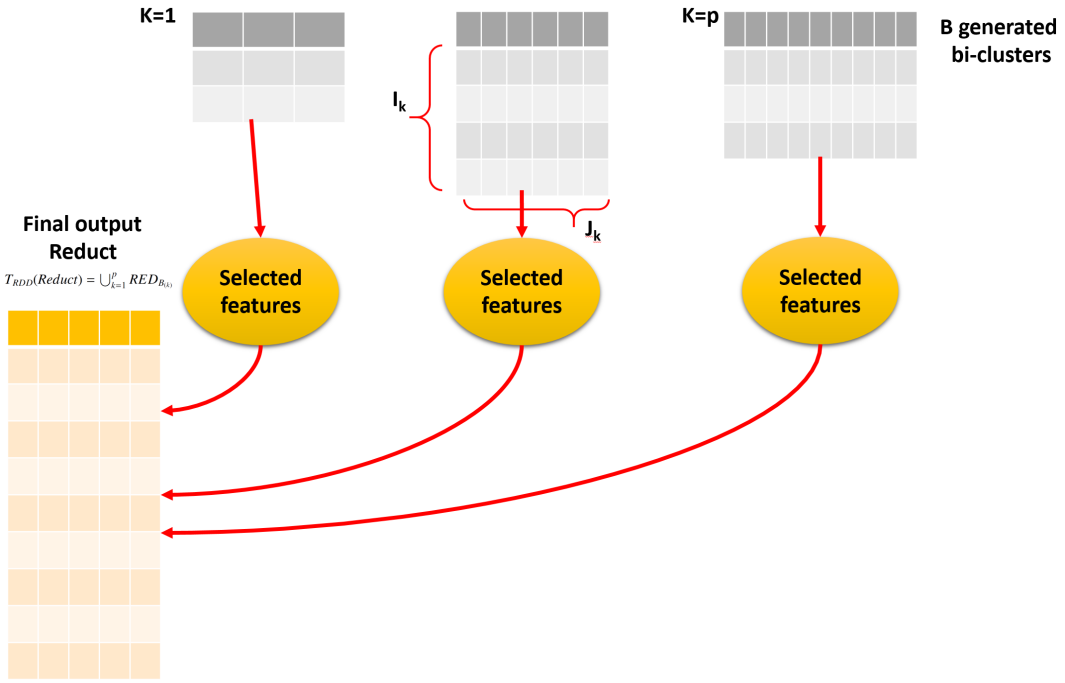

Figure 4.3: A high level description of the distributed RST bi-clustering based method

Contribution 3.2: Key conclusions This contribution describes a work that was initiated in collaboration with the *Department of Computer Science and Engineering*, at Bangladesh University of Engineering and Technology; where three Master students were involved.

4.2.3 Dissemination of results

The technical description of context 1 was published at the *IEEE Big Data Conference 2017 (IEEE Big Data)* [DZBL17]. The approach was also published at the French reputable conference *La Conférence Extraction et Gestion des Connaissances (EGC) 2018* [DZBL18b], and at the EGC workshop *Fouille de Donneée Complexes (FDC) 2018* [DZBL18a]. An extension of this work was published in the *Knowledge and Information Systems Journal* [DZBL20]. The technical description of context 3 (the hashing contribution) was published and presented at the *IEEE Big Data Conference 2018 (IEEE Big Data)* [DZB+18]. The paper was selected to be submitted as an extended version and was publsihed in the *Fundamenta Informaticae Journal* [CDZ21].

4.3 Privacy preservation for semantically enriched mobility data using granular computation and federated learning

This section outlines the initial research that was conducted as part of my contributions to the MASTER project (MSCA-RISE, described in Chapter 2 – Table 2.6). This section also covers ongoing work that is part of a PhD thesis that I am currently co-supervising (Chapter 3 – Section 3.2.2).

4.3.1 Context and motivation

The pervasiveness of smartphones and various connected sensors and wearables is leading to a wide collection of data including positioning and trajectories, reflecting the user's mobility. This movement data can be enriched with other semantic data that co-occur in space and time. Mobility data analysis and mining is of paramount interest for several applications such as traffic engineering, urban planning and environmental studies. Nonetheless, individual trajectories are highly sensitive information [DMHVB13], and their use is restricted [Dam14]. Eventually, the risks are exacerbated in the context of semantically enriched trajectories, that encompass more information on the individual habits, and surrounding context. Therefore, there is a need to develop solutions that could allow analyzing mobility and trajectories data while preserving the privacy of individuals. Eventually, individual privacy breach has become a paramount question today in society with all the collected data associated with high performance computing capacities and large-scale storage in the cloud.

Semantic trajectories are characterized by their different aspects or dimensions, and their complex nature [BRdA+14]. When it comes to trajectory representation, the same semantic trajectory can be represented with respect to different aspects [RBT+21]. As an example, a raw trajectory can be represented as a sequence of stops and moves, or as a sequence of transportation means, or as a sequence of weather conditions, of activities performed during the movement, and so on. This led to the *"Multiple Aspect Trajectory"* (MAT) concept [MBA+19]. The representation of MAT, called the MASTER model, addresses the issue raised by [FAB16], in which different aspects were taken into account independently. The model permits the representation of a trajectory in terms of space, time, and a variety of other aspects. Figure 4.4 depicts a multiple aspect trajectory with aspects that differ along the trajectory, as defined by [MBA+19].

In the rapidly evolving field of research on privacy preservation, granular computation and federated learning are emerging as highly promising approaches. In the contributions that will be discussed below, how granular computation and federated learning can be used to ensure privacy preservation will be elucidated in the domain of mobility mining.

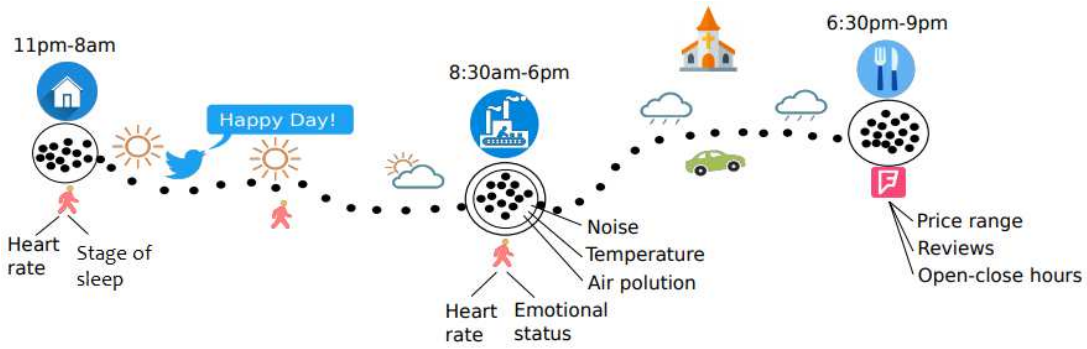

Figure 4.4: Example of a multiple aspect trajectory

4.3.2 Contributions

Contribution 1: Towards a granular computing framework for multiple aspect trajectory representation

My aim is to introduce a fresh research direction that involves the application of Granular Computing (GrC) formal settings (e.g., rough set theory, fuzzy set theory) to address the challenges of multiple aspect trajectory representation and privacy preservation.

A granular representation The concept of "multiple aspect" trajectories can be modeled using the hierarchical structure of granular computing. Data granules are formal entities that play a crucial role in organizing data and capturing relationship knowledge. The use of data granules can help to detect hidden patterns, as well as increase the logicality, systematicity, and efficiency of decision-making [ZTY19, CMMA+18].

A granular representation of multiple aspect trajectories is a complete entity with multiple dimensions, such as space, time, weather, transportation means, emotion status, and activity. As all of these dimensions are related to the spatial and temporal scales, then the granular representation of multiple aspect trajectories should consider the spatiotemporal aspect. This can be guaranteed by integrating the time and space dimensions, as attributes of data granules, into a single systematic model.

Figure 4.5 (inspired by [YRSF20]) represents the multiple aspect trajectory data granule structure. Every considered data granule $Mat-Gr$ is represented via a layer $Mat-Lay_l$, where $l \in \{1,\ldots,L\}$, and L is the maximum number of layers for the different considered dimensions. With respect to this structure, every $Mat-Gr$ can be defined in different scales; denoted as $Mat-Gr_l$. The data granules in each upper scale $Mat-Gr_l$ are transformed into those in the lower scale $Mat-Gr_{l+1}$ using a granulation criteria (rule) GCr_i, where $i \in \{1,\ldots,L-1\}$. The data granules decrease (increase) as the scale decreases (increases). A correlation, representing the connection between the different aspects of the multiple aspect trajectory, is reflected by the different edges between the various $Mat-Gr_l$. All $Mat-Gr_l$

are connected between each other as well as between their corresponding granules.

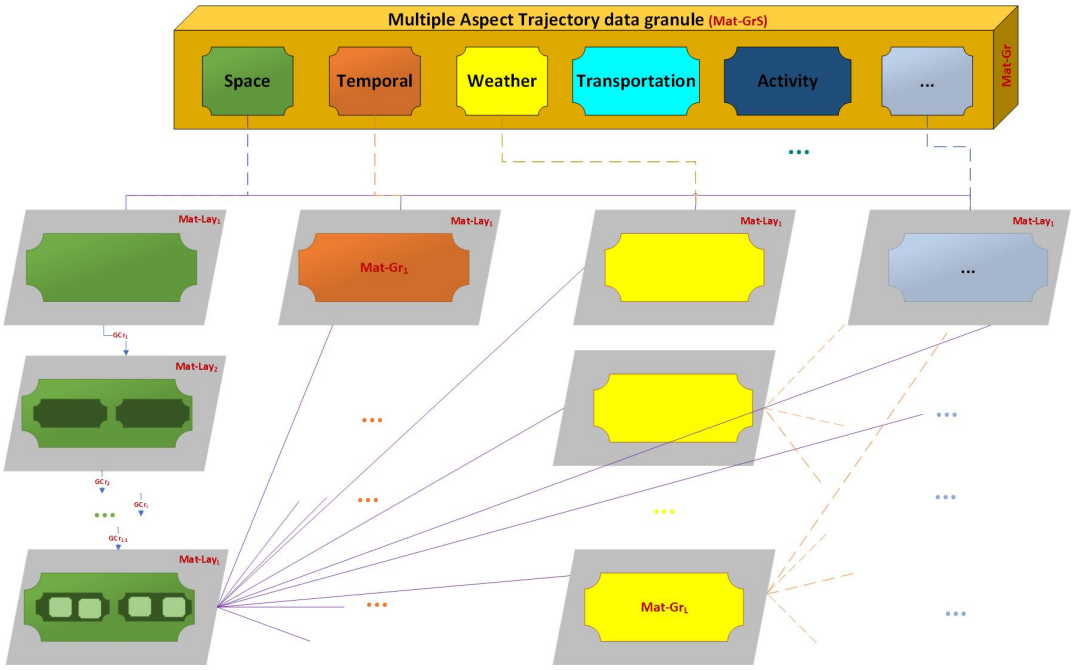

Figure 4.5: A granular representation of multiple aspect trajectories

Example of a granular representation of a multiple aspect trajectory. Let us suppose that during one stop the object is moving on foot and the weather condition changes from sunny to rainy. Figure 4.6 represents a simplified view of a type-2 granular structure of such representation. The edges represent the different correlations between the different granules; i.e., temporal-spacial, temporal-activity, space-weather, ..., and space-temporal-weather-activity.

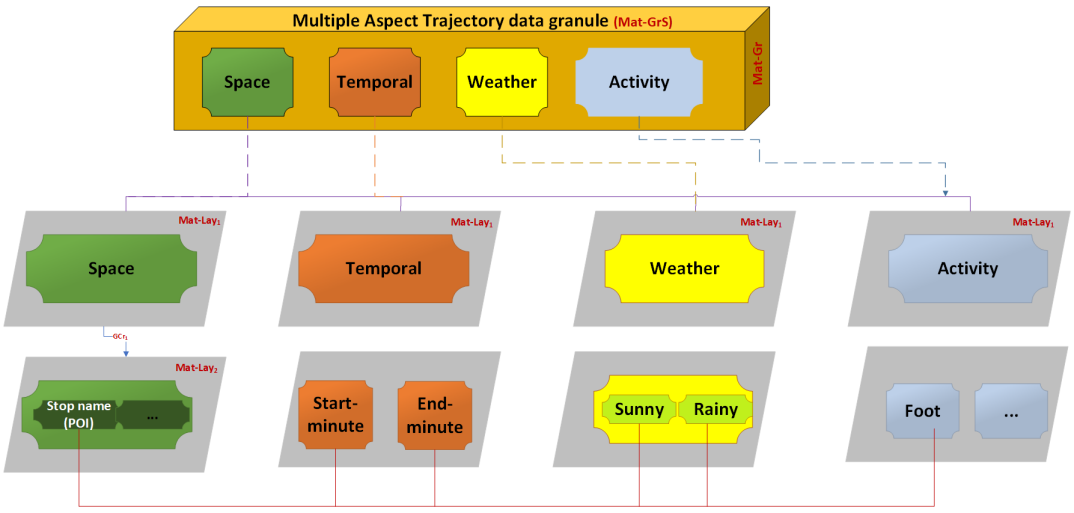

Figure 4.6: Example of a granular representation of a multiple aspect trajectory

With respect to this granular structure, we can define the following roots:

- At POI_x, Weather = Sunny during Start-time = t_v and End-time = t_w, and Activity = Foot
- At POI_y, Weather = Rainy during Start-time = t_i and End-time = t_j, and Activity = Foot

With respect to the need of representing different aspects at a time, sub-granular structures can be extracted from Figure 4.5. This is to achieve landscape law mining at a specific scale for different granules.

Example of reasoning. Weather as an example can be considered as a localized or a particular view or aspect of the MAT presented in Figure 4.6. Sunny and rainy are examples of granules of the sub-layer ($Mat - Lay_2$) of the weather granular layer ($Mat - Lay_1$). Data granules in $Mat - Lay_2$, for instance, can be viewed as an information table, with a binary values' representation, with respect to space and time. Binary values represent the presence or absence of certain aspects' conditions in relation to space and time. By considering the binary representation of the granules, it becomes possible to analyze the spatial and temporal patterns of the aspects within the trajectory. Also, by including information about the POI, it becomes feasible to explore the relationships between aspects, space, time, and specific points of interest within the trajectory.

Suppose you have a dataset that consists of different trajectories recorded over time. Each trajectory represents a person's movement throughout a day, and it is described by a set of binary variables: sunny (1 if it is sunny, 0 if it is not), windy (1 if it is windy, 0 if it is not), foot (1 if the person is traveling on foot, 0 if not), and car (1 if the person is traveling by car, 0 if not). If we consider Rough Set Theory as a GrC theory, the concept of an attribute set and its associated decision set is important. An attribute set is a combination of attributes, and the decision set represents the corresponding outcomes or classes. In this case, the attributes are sunny, windy, foot, and car, and the decision set could be the presence or absence of specific points of interest (POIs) within the trajectory. For example, let's consider a POI called "Park" represented by a binary variable Park (1 if the trajectory includes a visit to the park, 0 if not). Rough set theory can help identifying the dependencies between the aspects (sunny, windy, foot, car) and the decision attribute (Park). By calculating the lower and upper approximations, we can determine which attribute combinations are necessary or sufficient to determine the presence or absence of the Park; with respect to space and time. Indeed, rough set theory allows the identification of trajectories that exhibit unusual or inconsistent behavior with respect to the presence or absence of the park. Objects that do not conform to the lower or upper approximations (by checking the boundary region) can be considered as outliers, providing insights into unique patterns or anomalies within the dataset.

The privacy challenge In any representation of trajectory data, the deductive route from non-sensitive to sensitive features can pose privacy risks. The sensitive features can be

defined as the set of features that should be hidden using algorithms in such a way that they cannot be reasonably approximated using the set of non-sensitive features. This is quite a challenge since the non-sensitive features may contain concepts on which sensitive features inferentially depend; these are known as *"quasi-identifiers"*. For example, consider a dataset containing the trajectories of taxi rides in a city:

- *Non-sensitive feature:* Start and end locations of the taxi rides.
- *Sensitive feature:* Political rally attendance.

While the start and end locations may not directly reveal an individual's political affiliations or rally attendance, the deduction of this sensitive feature becomes possible when correlated with external events or publicly available information. By combining the trajectory data with the knowledge of rally locations and timings, an attacker could infer whether a particular individual attended political rallies.

To protect privacy in this scenario, identifying and handling quasi-identifiers is crucial. Quasi-identifiers in trajectory data could include: *Time of taxi rides* or *Frequency of taxi rides in specific areas.* These quasi-identifiers indirectly relate to the sensitive feature of political rally attendance. By analyzing the temporal patterns of taxi rides and identifying frequent rides in the vicinity of rally locations during specific time periods, an attacker could deduce an individual's political involvement. To cut the deductive route to any potential privacy leak, these quasi-identifiers must be appropriately hidden. However, they must be uncovered first. In this concern, rough set theory as a GrC example, seemed to be one among the right tools to be used to undermine the deductive route from non-sensitive to sensitive features by modelling and analyzing the dependency "non-sensitive $\longrightarrow$ sensitive" [BCP+10]. Rough set theory provides numerical quantities that measure the degree of dependency of attributes. This study was presented in [WSNKK20], where authors have learned the quasi-identifiers, computed a granulation of the information system that maximizes the distribution of sensitive features in each granule, and masked the deductive route from non-sensitive to sensitive attributes.

Granular computing methods offer plenty of opportunities to handle the challenges tied to multiple aspect trajectory representation and privacy preservation. This new research area that I propose will present a considerable addendum to the mobility data management, analytic and privacy communities and fields.

Contribution 2: A privacy-preserving solution for semantically enriched mobility data using federated learning

The research work discussed in this section is in its early stages as it forms part of a PhD thesis, commenced in October 2022 (Chapter 3, Section 3.2.2), that I am currently co-supervising. There are several learning tasks to consider such as trajectory classification, prediction, clustering, mobility pattern mining, local / global trajectory analysis and statistics. However,

adequate public mobility data sets and benchmarks do not exist. To be able to evaluate the proposed methods, our first investigation focuses on the generation of synthetic "rich trajectories" that mimic the real (private) ones. Specifically, we first aim to investigate, study, and apply a Federated Learning (FL) approach in the context of privacy preserving semantically enriched mobility mining, which to the best of our knowledge has not been fully investigated yet. Applying the FL paradigm to semantically rich mobility data is a very challenging task due to the complexity of the enriched trajectories, and their continuous evolution in both time and space. Thus, rethinking the FL algorithms and the overall orchestration architecture is highly required as we have emphasized in [CDRZA20].

Problem statement Synthetic data generation based on generative adversarial networks (GAN) [AMB21] has gained attention as a promising solution to address privacy concerns. GAN trajectory generation methods have been used to protect the geo-privacy of trajectory data, with recent studies showing that synthetic records can achieve comparable performance to real data in multiple tasks [LCA18, OSRY18, RGKH20, WLLY21].

Despite the potential of GAN trajectory generation, most methods rely on a large dataset of true trajectories collected from individual users in a centralized manner. However, the assumption of sharing personal sensitive data does not hold in many real-world scenarios. In fact many users do not want to share their travel information with untrusted external entities. As such, finding alternative ways to generate synthetic data without the need for centralized data collection is crucial. This requires the development of new techniques that can generate realistic synthetic data while preserving the privacy of individual users. Moreover, there is a need to ensure that the generated synthetic data is representative of the true trajectories and preserves the statistical properties of the original data. This is necessary to ensure that the synthetic data can be used effectively in various applications, such as route planning and traffic analysis.

Description As an initial contribution, we present a decentralized approach for training TimeGAN [YJVdS19] to generate synthetic trajectories, as depicted in Figure 4.7. To ensure privacy of mobility data, we leverage federated learning [MMR+17] as a privacy-preserving machine learning technique. This approach enables participants to collectively build a model without revealing their individual data. TimeGAN is specifically designed for multivariate time series, which makes it an ideal option for generating synthetic trajectories; semantically enriched.

Our trajectory generation framework follows a similar processing method to existing solutions. However, unlike these previous methods, we do not collect the data first. Instead, each individual user runs their own TimeGAN model on their local data to generate new synthetic trajectories. The user then sends their model weights to a centralized server after a set number of iterations. The server aggregates the local models and sends back new global

parameters to each user. This approach ensures that individual user data is kept private and secure, while still allowing for the creation of realistic synthetic trajectories.

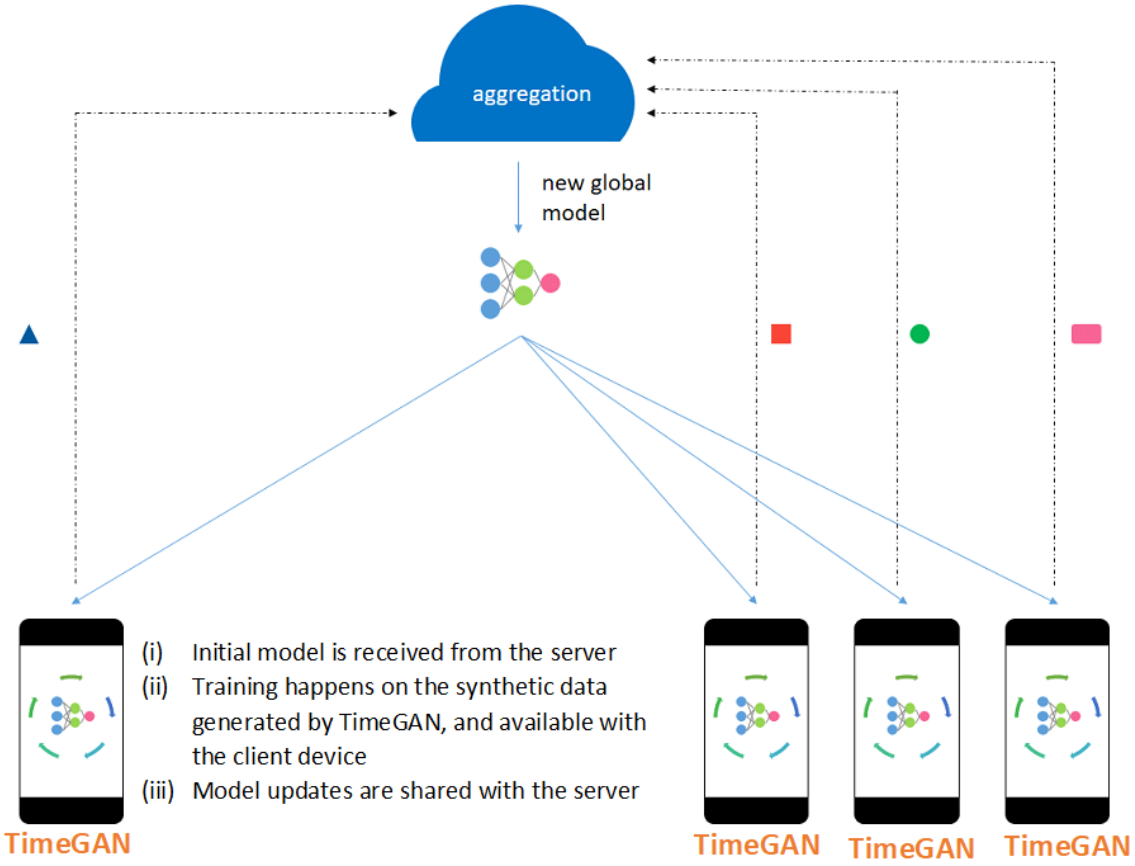

Figure 4.7: An overview of the FL-TimeGAN proposed solution

Key conclusions The first technical contribution of this work is currently under development and is a work in progress.

4.3.3 Dissemination of results

The description of contribution 1 (Section 4.3.2) was published at the *4th International Symposium on Rough Sets: Theory and Applications (RSTA'22)*, part of the *17th Conference on Computer Science and Intelligence Systems (FedCSIS)* [DB22]. The motivation behind investigating privacy preservation in semantically enriched mobility data, in contribution 2 (Section 4.3.2), was published in the *1st Workshop on Flexible Resource and Application Management on the Edge (FRAME'2021)*, which is in association with the *30th International ACM Symposium on High-Performance, Parallel and Distributed Computing (HPDC)* [CDRZA20]. The first technical contribution is a work in progress and hence has not been disseminated yet.

4.4 Conclusion

This chapter provides an overview of the various contributions and work in progress made towards the development of novel approaches for knowledge discovery and privacy preservation using granular computation and federated learning. The next chapter will describe the second research direction of this Habilitation Thesis.

Chapter 5

New approaches for knowledge discovery using biological computation and deep learning

5.1 Introduction

The objective of this chapter is to introduce some of the work conducted in the second research avenue pursued in this Habilitation Thesis. It delves into the realm of biological computation and deep learning, exploring their applications in knowledge discovery.

The chapter is organized as follows: Section 5.2 primarily focuses on the advancements made in the field of artificial immune systems. Section 5.3 explores novel evolutionary approaches for the detection of Android malware. Section 5.4 introduces an evolutionary approach designed for the Bring Your Own Device (BYOD) scenario. The contribution made in solving the wireless sensor network deployment problem from an evolutionary perspective is described in Section 5.5. In Section 5.6, a new multi-objective multi-agent deep interactive reinforcement learning approach is presented. Finally, the conclusion is presented in Section 5.7.

5.2 Advances in artificial immune systems and some theoretical work

5.2.1 Context and motivation

Most of the advanced bio-inspired algorithms face challenges when dealing with large amounts of data. To address this issue, researchers have developed a collection of parallel evolutionary algorithms [AT02, Alb06, Alb05]. The objective of this section is to enhance the application of nature-inspired algorithms to big data by adapting the Dendritic Cell Algorithm (DCA) in a practical manner. Although the DCA has emerged as a promising approach, its practical implementation has been limited to problems of moderate size in existing literature. Another aspect explored in this section is the utilization of artificial immune systems and evolutionary algorithms in Astronomy and Geoscience, demonstrating several successful applications in these fields. Additionally, the section discusses the convergence between

biological computation and computational biology.

5.2.2 Contributions

Contribution 1: A scalable dendritic cell algorithm

The first contribution highlights the developed Sp-DCA solution, which is a bio-inspired approach that can effectively perform big data classification.

Problem statement The practical application of the DCA has been limited to problems of moderate size. This limitation arises from the algorithm's use of an antigen multiplier (m) for classification purposes. Each data instance (x_i) is replicated (m) times, resulting in the generation of $[x_i] * m$ antigens. Considering the size of the dataset (N), a total of $[x_i] * m * N$ antigens would be created from the entire input dataset. However, due to hardware and memory constraints, generating all these antigen copies becomes impractical, given the large number of data instances. To overcome these limitations and broaden the applicability of the DCA in real-world scenarios, the development of a distributed version of the DCA was deemed essential.

Description A high level description of Sp-DCA is presented in Figure 5.1, whereas a description of the approach is given in what follows.

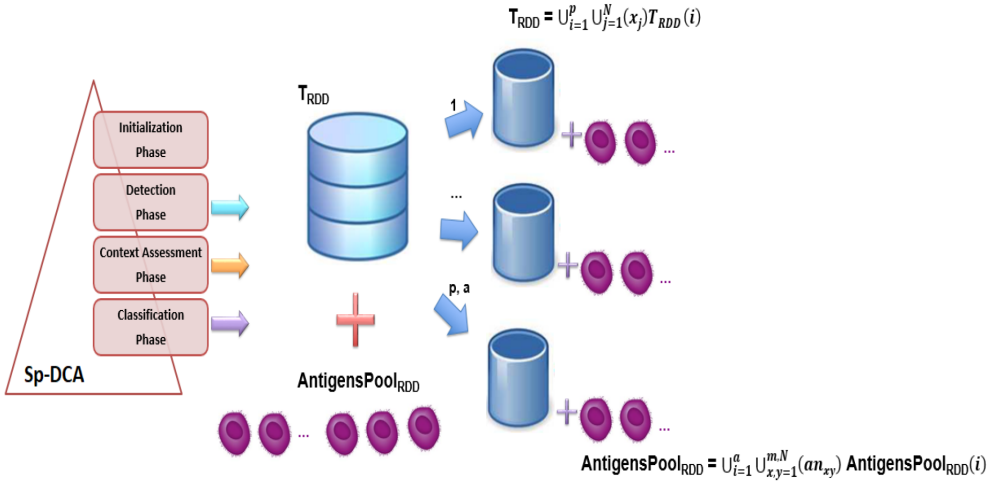

Figure 5.1: A high level description of the distributed dendritic cell algorithm

- The antigen data set of size N is defined as T_{RDD}, where universe $U = \{x_1, \ldots, x_N\}$ is the set of antigen identifiers, the attribute set $C = \{c_1, \ldots, c_V\}$ contains every single feature of the T_{RDD} and the decision attribute D corresponds to the class label of each T_{RDD} sample. As Sp-DCA is based on the standard DCA concepts, and since DCA is applied to binary classification problems; then the decision attribute, D, of Sp-DCA is defined as: $D = \{d_{normal}, d_{anomalous}\}$. The universe set U presents the pool from

where the antigens will be multiplied by an antigen multiplier m generating a pool of antigens: $AntigensPool = \{an_{1_i}, \ldots, an_{1_m}, \ldots, an_{N_i}, \ldots, an_{N_m}\}$.

- In order to make Sp-DCA scalable with the high number of both training data and antigens, Sp-DCA partitions the given T_{RDD} into p data blocks based on splits from the universe set U. Indeed, Sp-DCA creates an RDD from the generated antigens pool, $AntigensPool_{RDD}$, and splits it into a a number of disjoint subsets. This leads to the following: $T_{RDD} = \bigcup_{i=1}^{p} \bigcup_{j=1}^{N} (x_j) T_{RDD_{(i)}}$ and $AntigensPool_{RDD} = \bigcup_{i=1}^{a} \bigcup_{x,y=1}^{m,N} (an_{x,y}) AntigensPool_{RDD_{(i)}}$.
- To ensure scalability, the distributed algorithm will be applied to every single $T_{RDD_{(i)}}$ and $AntigensPool_{RDD_{(i)}}$ that at the end all the intermediate results will be gathered from the different p and a partitions.
- All of the DCA phases described in Appendix 8 are redesigned and developed in a distributed fashion as described in [Dag18a, Dag18b].

Key conclusions Developing a distributed schema based on *MapReduce* for the DCA motivates the global purposes of this contribution which are (i) to enable the DCA to deal with big data classification problems (ii) to illustrate the scalability of the proposed schema and (iii) to analyse the behaviour of the proposed solution within a distributed environment particularly in terms of classification performance.

Results in [Dag18a, Dag18b] revealed that the Spark paradigm has offered an efficient environment to parallelize the functioning of the DCA allowing it to overcome its memory and runtime restrictions. In terms of classification performance, results showed that Sp-DCA holds its sensitivity characteristic. Based on this evaluation, it is concluded that the classification performance bottleneck of the DCA is not addressed by the use of a distributed implementation design; as expected.

Contribution 2: When evolutionary computing meets astro- and geo-informatics

The second contribution presents a theoretical study of different artificial immune systems and evolutionary algorithms in Astronomy and Geoscience. The contribution was made part of the BIG-SKY-EARTH COST Action project (Chapter 2, Table 2.9).

Description In this contribution, and in close collaboration with the *Department of Computer Science and Engineering, Saints Cyril and and Methodius University*, we have conducted a thorough study of artificial immune systems and evolutionary algorithms to present a palette of techniques inspired by nature and which tend to solve the complex optimization problems encountered in astronomy and geoscience. We discussed the fact that evolutionary computing brings plenty of benefits in facing optimization challenges. As it was outlined in [Fog97], generally they are conceptually simple, have a wide range of applications,

allow for an external knowledge to be incorporated, can be self-adaptive, and can be applied to complex problems. Moreover, they can be easily parallelized.

Key conclusions We gave an overview of various methods from evolutionary computation and provided some interesting examples of their applications to real-world problems in astronomy and geoscience hoping to motivate the reader in employing the algorithms to the problems they are facing. We highlighted that comparisons of the different algorithms can be made from many aspects, but typically the appropriateness of the method depends on the particular application and how the problem is formulated. We emphasised that it is better to first explore the previous similar experiences for the specific problem, before deciding which algorithm is worth trying.

Contribution 3: The convergence of biological computation and computational biology

The third contribution presents another theoretical investigation exploring a large pool of algorithms inspired by nature, and discusses the convergence of biological computation and computational biology.

Description Biological computation involves the design and development of algorithms and computational techniques inspired by natural biota. On the other hand, computational biology involves the development and application of computational techniques to study biological systems. Thus, biology and computer science can guide and benefit each other, resulting in improved understanding of biological processes and at the same time advancing the design of algorithms. However, integration and fruitful collaboration between biology and computer science is often very challenging, especially due to the cultural idiosyncrasies of these two communities. Despite notable advances in computational biology and biological computation over the last few decades, we believe that a more involved integration is required to comprehend and predict the complexities of biological systems, and improve the existing biologically based computational paradigms.

Key conclusions In [CDAB21], we aimed at highlighting how nature has inspired the development of various algorithms and techniques in computer science, and how computational techniques and mathematical modelling have helped to better understand various fields in biology. We identified existing gaps between biological computation and computational biology and felt the need for a more involved communication between them. We advocate for inspiring researchers in the intellectual adventure of working on the interface of biology and computational thinking, leading to a firm and broad foundation for a deeper integration between "wet" (lab-based) and "dry" research. Therefore, along with others [FH07, BL16, Nur08, Pri09, NBJ11, RC14], we are advocating for the need of developing advanced and appropriate computational tools to better understand and model life sciences,

as well as the need for a deeper integration of computer science in biology.

5.2.3 Dissemination of results

The technical description of Sp-DCA was published at the *Genetic and Evolutionary Computation Conference 2018* (GECCO'2018) [Dag18a]. The approach was, also, published at the *Swarm and Evolutionary Computation Journal* [Dag18b]. The work conducted in contribution 2 is part of the *Knowledge Discovery in Big Data from Astronomy and Earth Observation*[1] book in which our chapter was published [DM20]. As for contribution 3, it represents the outcome of an analysis that has been done in the 5^{th} Heidelberg Laureate Forum; specifically during a workshop[2] organized by myself and mentored by *Professor Stephen Smale (Fields Medal awardee)*. After the workshop, I set up a collaboration with two participants and contributors to the workshop, Pavel Avdeyev (PhD) from the *Department of Mathematics and Computational Biology Institute, The George Washington University* and Dr. Md. Shamsuzzoha Bayzid from the *Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology* who work on different areas in computational biology. This contribution was published in the *Artificial Intelligence Review Journal* [CDAB21].

5.3 New evolutionary and granular approaches for Android malware detection

Research presented in this section was conducted in the context of a co-supervision of a PhD student (Chapter 3 – Section 3.2.4).

5.3.1 Context and motivation

Mobile devices have become undoubtedly an inseparable component of one's daily life, allowing a quick and imminent access to different applications and services like allowing users to check emails, access online banking services, check social networks, etc. However, the rapid development of mobile Internet technologies led to major cyber threats; one of these is the presence of malware. By malware, we refer to any software that performs malicious actions, including information stealing, espionage, etc., and that inflicts harm on devices in multiple ways. While the diversity of malware is increasing, anti-virus scanners cannot fully fulfill the needs of protection, resulting in millions of hosts being attacked. Current commercial anti-virus tools might not provide efficient detection, especially when dealing with zero-day attacks. Also, the performance of the detection process heavily depends

[1] https://www.elsevier.com/books/knowledge-discovery-in-big-data-from-astronomy-and-earth-observation/skoda/978-0-12-819154-5

[2] https://www.heidelberg-laureate-forum.org/blog/experience-learn-and-share-at-the-heidelberg-laureate-forum/

on the base of examples of malware behaviors. Actually, malware developers usually use obfuscation techniques consisting in a set of transformations that make the code and/or its execution difficult to analyze by hindering both manual and automated inspections. These techniques allow the malware to escape the detection tools, and hence to be seen as a benign program. For this reason, other detection approaches have emerged in literature. Among these are approaches which create new malware and variants, of known malware relying, for the most part of them, on Evolutionary Algorithms (EAs). The generation process of those malware variants is essentially based on the extraction of frequent Application Programming Interface (API) call sequences of previously encountered malware programs. This allows building a base of fraudulent behaviors, also called malicious patterns. Despite the interesting performance of these approaches, most of them present some flaws: some of them are not fully automated as some tasks were done in a manual manner like labeling the used patterns, while some others are only proposed for a specific attack type, and most of them neither check the consistency of the produced detection rules nor the trustiness and reliability of the generated attacks, i.e., if they are certainty malicious or not.

In this work, we investigated new pathways for malware detection. In this concern, we have opted for evolutionary algorithms and granular computation to propose novel evolutionary based detection techniques which are capable of outperforming state-of-the-art methods while dealing with the above mentioned shortcomings.

5.3.2 Contributions

Contribution 1: Artificial Malware-based Detection

Problem statement State-of-the-art malware detection techniques achieve high reported detection performance on predefined data sets. However, most of them are less suitable for real-world deployment because requirements for malware detection need to be independent from the provided base of examples of malware behaviors. Several state-of-the-art techniques have proposed to extract frequent API call sequences from encountered malware programs using pattern mining techniques. These sequences build a base of fraudulent behaviors. Afterwards, API call sequences can be extracted from any program and based on these, the considered program behavior can be judged to be more similar to malware behaviors or benign-ware ones. The main remark that can be extracted from current detection techniques is that the efficiency of malware detection is very dependent on the base of examples of malware behaviors. More precisely, the provided or collected base of examples is static and therefore lacks diversity, which may negatively impact the detection rate.

Description The first proposed approach is a hybrid detection method, called "Artificial Malware-based Detection" (AMD), that makes use of not only extracted malware patterns but also artificially generated ones which are generated using an evolutionary *Genetic Algorithm* (GA). Our proposed AMD approach functions as follows:

1. *API call sequences extraction:* This first phase is responsible mainly for extracting the API call sequences from the collection of normal and malicious applications to transmit them afterwards to the next phase.
2. *Patterns construction:* Through this phase, the process of the patterns construction is subdivided into three main sub-steps namely the extraction of the frequent API call sequences, the generation of the artificial patterns and the patterns gathering process. In the first sub-step, the frequent sets of API call sequences, referred to as frequent item sets (also called patterns), are extracted. Among these, a selection is performed to keep a set of the unique patterns. In the second sub-step, a database of artificially generated malware patterns is created using the set of the selected patterns. This is achieved via the use of a Genetic Algorithm, aiming to diversify the base of malware examples with artificial malicious patterns in order to maximize the detection rate. The third sub-step will mainly gather the results of the two previous sub-steps and thus, has two outputs: in one hand it generates a collection of benign patterns, and on the other hand it offers an enriched collection of malicious patterns, i.e., the selected malware patterns and the artificially generated ones.
3. *The detection phase:* In this phase, unknown new apps are classified. Specifically, by using the collection of the generated patterns, malicious programs will be detected among the new apps.

Key conclusions Our results reported in [JDBS20] showed promising results. Through generating both the malware pattern set and the benign pattern set, our AMD proposed approach outperforms the state-of-the-art well-known detection approaches with better detection accuracy rates. Moreover, by analyzing the false positive and false negative values, we noticed that our AMD approach was able to achieve better results when including the artificially generated patterns. Based on the conducted evaluations and the promising obtained results, we have shown that our AMD approach is an interesting malware detection approach that is indeed able to detect obfuscated malware thanks to the use of the artificially generated patterns.

Contribution 2: Bi-Level Malware Detection

Problem statement The previously proposed AMD approach presents some limitations: (i) the detection rule is pre-specified independently of the generated malware patterns; and thus the rule definition and the classification tasks are done separately without any interaction, and (ii) in AMD, the artificial malware patterns are generated in a global ad-hoc manner based on their similarities to real malware and benign patterns from the bases of examples, which may increase the number of false positives. These limitations will be overcame in this second contribution.

Description The second contribution is a bi-level detection method, called "Bi-Level Malware Detection" (BMD), where we suggested considering the detection rules generation process as a BLOP (Bi-Level Optimization Problem), where a lower-level optimization task is embedded within the upper-level one.

In the bi-level formulation of BMD, the lower-level problem allows to find new malicious artificial patterns. The evaluation of a detection rule is based on its accuracy using both the base of examples (input) and also the artificial malicious patterns generated by the lower-level problem. We aimed to maximize the detection accuracy rate of the rules. The follower (lower-level) uses patterns from the base of examples, i.e., the set of malicious and benign patterns, to generate artificial malicious patterns. A GA is used in order to perform the generation process of artificial malicious patterns that maximizes not only the number of new artificial malicious patterns but also the number of generated malicious patterns that are not detected by the leader (detection rules). The upper-level keeps exchanging solutions with the lower-level, i.e., the upper-level sends detection rules to the lower-level and the lower-level sends the generated artificial malicious patterns to the upper-level, until a stopping criterion is met (e.g., number of iterations). Within these exchanges, the detection rules generated by a genetic programming algorithm (GP) are improved from one iteration to another as they are capable of detecting the new generated malicious patterns. On the other side, within the lower-level, the generated malicious artificial patterns are improved from one iteration to another that they can escape being detected by the produced detection rules which are sent by the upper-level. At the end of these exchanges, the best detection rules present the final output produced by our BMD approach.

Key conclusions Based on our results reported in [JDBS22], our BMD approach revealed its robustness to new attacks and hence can be considered as an interesting malware detection approach thanks to the use of the bi-level approach.

Contribution 3: Rough-Set Based Bi-level Malware Detection

Problem statement Further investigating the BMD approach allowed us to highlight the following limitations: (i) the detection rule is prespecified depending on the generated malware patterns; and thus the rule generation task made by the GP may cause inconsistency within the detection rules specifically when applying the crossover operator, and (ii) BMD offers the user a two-way crisp decision labeling (malicious/normal). These shortcomings will be overcame in this third PhD Thesis contribution.

Description We developed a new effective and efficient malware detection method dubbed "Rough-Set Based Bi-level Malware Detection" (RS-BMD) that improves its detection rate thanks to the integration of two key components: (i) BLOP, and (ii) Rough Set Theory. In RS-BMD, the malware detection rule generation process is formalized as a BLOP; as the case of BMD [JDBS22]. RST is integrated in both levels of the RS-BMD BLOP schema. First it

is coupled with the used GA, within the lower-level, for malicious patterns evolving, and second, within the upper-level, it is coupled with GP for malware detection rules induction. Coupling RST with EAs is a promising way able to induce efficient and reliable decision rules even from inconsistent information. By using the RST fundamentals, RS-BMD will be able to evaluate the real nature of the artificially generated malicious patterns (within the lower-level), to only keep the certain malicious ones, to check the reliability of the generated detection rules (within the upper-level), and to only keep the most efficient and consistent ones, and to classify in a three-way decision fashion the detection rules as: accept which refers to the certain rules for acceptance (i.e., benign apps), abstain which refers to the possible rules for indecision or delayed decision, and reject which refers to the certain rules for rejection (i.e., malicious apps). This will not only allow to deal with the artificial malicious patterns' and the detection rules' inconsistencies but also to provide a more coherent way for decision making.

Key conclusions Our malware detection method has shown its merits when compared to several state-of-the art malware detection techniques. With respect to the obtained results, partially reported in [JCDBBS21], we could deduce that when we assure a continuous variability to our base of examples, guaranteed by the BLOP component, and by injecting the instances from the set of the certain generated malicious patterns, guaranteed by the RST component, we guarantee a better detection of malware. We also concluded that RST helped RS-BMD keeping a better set of detection rules by removing the inconsistent ones, and hence could confirm that the Rough set component helped keeping the malicious patterns of better quality. Those generated malicious patterns will be fed to the upper-level and consequently improve the quality of the generated detection rules produced by this level.

As for the benefits of using a three-way decision making fashion, we analyzed the generated number of possible rules and their usefulness for the end user. The obtained set of possible rules will grant the user to make an extra choice about the fate of an app which is not possible with other methods. In fact, in other approaches, this app would be more likely to be denied from accessing the system and consequently deprive the user from using it.

5.3.3 Dissemination of results

A first journal paper revealing the first contribution was published in the *Computers & Security Journal* [JDBS20]. The second contribution was also published in *Computers & Security Journal* [JDBS22]. A subset of the third contribution was published in the proceedings of the *28th International Conference on Neural Information Processing (ICONIP 2021)* [JCDBBS21]. The full contribution is submitted to the *Cognitive Computation Journal*.

5.4 An evolutionary approach for the BYOD context

Research presented in this section was conducted *jointly with a PhD student and other collaborators from the Department of Computer Architecture and Computer Technology, ETSIIT and CITIC, University of Granada, and from the Department of Computer Engineering, ESI, University of Cádiz, Spain.* This work makes use of the data gathered from the MUSES project (Chapter 2, Table 2.10).

5.4.1 Context and motivation

The fast pace of introduction of new technologies has led modern computing to undergo several outstanding transitions in a short period of time. Modern computing has moved over time to smaller, more reliable, and faster high-tech devices such as smartphones, laptops, and tablets. The use of these technologies in several forms is progressing, and has led to the form of use known as the *"Bring Your Own Device"* (BYOD) concept or philosophy.

In the corporate world, the BYOD practice refers to allowing employees to use their personal laptops, smartphones, tablets, and other mobile devices for work-related tasks, but not necessarily while being in the workplace. This has many advantages [Sin12]; among them we mention saving costs — since the company saves money on high-priced devices that it would normally be required to purchase for their employees —, and increasing flexibility and worker productivity as employees will not be asked to haul around multiple devices to satisfy both their personal and work needs, and having everything they need in one device anytime and anywhere. Other advantages are tied to the increase of worker satisfaction, attracting the best candidates, and the increase of engagement in the workplace and after hours [DD19].

Despite the significant advantages of this policy, BYOD exposes companies to security threats such as leak of confidential data and unauthorized access. The security policy rules needed to control this are usually fixed and proposed by the Chief Security Officer (CSO).

5.4.2 Contribution

Problem statement There is a big disadvantage concerning security since potentially unsecured devices from unaware users might interact with important company assets [ALL+18]. The main issue in the adaptation to a BYOD scenario is to obtain a high level of security, while maintaining user privacy [MVH12]. Clearly, the uncontrolled access to internal networks by the personal devices, for which companies have control limits due to privacy concerns [MVH12], exposes the companies to security risks such as data leakage, surveillance, and many other possible threats [Len12]. These threats have become the companies main security concern, which makes a challenge to assure a compromise between pushing personal devices towards professional use while coping with their own stringent and complex security requirements [ALL+18]. This trend is inevitable as enterprises are faced with questions of whether and how to manage this situation [Tho12].

To this end, the Corporate Security Policies (CSPs) [Kae03], approved by the company's CSO are the core at the identification of threats and the construction of a set of security rules. The description of these jobs include protecting company assets by defining permissions to be considered for every different performed action inside or outside the company's work space, and eventually coming from the employees personal devices. Nonetheless, CSOs build the set of CSPs based on their expertise, and as such, they have the limitation of not knowing every possible combination of events that might lead to a dangerous situation. Therefore, the contribution discussed in this part is in the form of a novel technique for extracting inference rules from past behaviour instances that might help the CSO in a BYOD environment in the definition and refinement of security policies that, eventually, classify an upcoming event or user action as permitted or not permitted.

Description The objective of this contribution is to obtain a way of correctly classifying the most incoming events by avoiding mostly the false positives. By false positives, we mean that events that must not be permitted should never receive an "OK", that is, "GRANTED".

We main to create a reliable rule set which is able to cover every new situation that may be a threat; allowing the system to go beyond the limited set of known pre-defined rules. In order to have the space of possible policy rules to be as wide as possible, it is necessary to have a technique that explores the rule space efficiently and with the least assumptions about rule structure. In this concern, the evolutionary GP approach was applied to discover novel and interesting knowledge, and rules from large amounts of data [Fre02]. The formalization of our work can be defined as follows:

- The assigned classes, or leaves of the tree, would be "allow" or "deny", acting over a certain incoming event; whilst the nodes are the conditions that have to be met to apply the action.
- GP is used to generate these classification trees, optimising an objective function called *fitness* which is defined as the accuracy of a rule or set of rules, along with the classification error. But since there are other metrics that influence in "how good" a rule or a set of rules is, such as the depth of the created tree, the number of nodes it has, or the obtained false positives, these were also covered by the defined fitness.
- The last step is to present the GP generated rules to the CSO of the company and tune the algorithm according to his/her decision; i.e., either to include or not the set of rules in the main security policy.

Key conclusions In [dlCGSD+20], we could establish a proof of concept to demonstrate the viability, and applicability of the GP approach to the BYOD security context for automatic rule discovery; hence, helping the CSO of a company discovering dangerous events.

5.4.3 Dissemination of results

The fruit of this work was published in the *International Conference on the Applications of Evolutionary Computation* [dlCGSD+20].

5.5 An evolutionary approach for wireless sensor network deployment

This work was conducted in the context of a co-supervision of a Master student (Chapter 3, Section 3.2.7).

5.5.1 Context and motivation

Sensor networks are described as a large number of spatially distributed autonomous sensor devices [Mar10]. Generally, nodes are networked through wireless, i.e., Wireless Sensor Networks (WSNs), in order to gather and analyse environmental information and collect the data in a central location. Owing the capabilities to monitor large areas, access remote places and react in real-time, WSNs have found their way into a wide variety of applications and systems such as environment monitoring, habitat monitoring, home automation, industrial automation, water monitoring and so on. Keeping in view that the deployment of a WSN application can be influenced by many factors, its implementation introduces several challenges. In this contribution, we focused on one of the fundamental issues in WSN designing which is the *deployment problem*. WSN deployment encompasses the determination of positions for sensor nodes in order to achieve intended coverage, connectivity and energy efficiency while keeping the number of nodes as minimum as possible [BB08]. Being similar to many real-world design problems related to engineering, the deployment problem is inherently characterized by the presence of multiple objectives which may or not conflict with each other. In this work, we proposed a WSN deployment optimization model that considers more than three objectives to optimize.

5.5.2 Contribution

Problem statement Based on recent surveys [INA+16, FLY+16] several problems related to WSNs could be modelled as optimization problems. It has been shown that these problems are multi-objective or many-objective by nature as usually different conflicting criteria should be considered simultaneously to come up with better solutions. Several works have demonstrated the benefits of considering two or three objectives instead of a single one. However, since the deployment of real-world WSNs encompasses more than three objectives, a multi-objective optimization may harm other deployment criteria which are conflicting with the selected ones. Therefore, it seemed crucial to extent this area of research and explore the modelling and the resolution of WSN deployment in a many-objective fashion.

Description The proposed deployment model, named "$WSN-\theta-DEA$", includes the optimization of the *number of clusters*, the *number of High-Signal Range (HSR) nodes*, the *number of Low-Signal Range (LSR) nodes*, the *coverage*, the *connectivity*, the *energy consumption* and the *throughput* which are competing with each other. The "$\theta-DEA$" was chosen to optimize the proposed model. Accordingly, the adaptation of the "$\theta-DEA$" to our model was considered. The choice of "$\theta-DEA$" was mainly based on its ability to handle problems with an important number of objectives, and especially, because "$\theta-DEA$" evaluates solutions while insuring both convergence and diversity.

Key conclusions Our results reported in [BACDBBS22] revealed the effectiveness of "$WSN-\theta-DEA$" when compared to mono- and multi-objective models, in terms of finding solutions with deployment schemes improving the overall compromise between objectives. Specifically, our proposed many-objective model is very competitive regarding the overall trade-off between the considered objectives. The model could offer solutions that are too close to solutions obtained by low-dimensions models. By using "$WSN-\theta-DEA$", the decision maker can optimize a large set of objectives together instead of improving certain criteria to the detriment of others.

5.5.3 Dissemination of results

A journal paper revealing this contribution was published in the *Evolutionary Intelligence Journal* [BACDBBS22].

5.6 A new multi-objective multi-agent deep interactive reinforcement learning approach

The research work discussed in this section is in its early stages as it forms part of a PhD thesis, commenced in October 2022 (Chapter 3, Section 3.2.1), that I am currently co-supervising.

5.6.1 Context and motivation

Multiomics data, characterized by their high dimensionality and complexity, pose a challenge in terms of handling and extracting valuable information. Integrating multiple omics data types allows for a comprehensive understanding of biological systems; however, it necessitates considering all data sources together. By treating multiomics data holistically, researchers can uncover intricate interactions and dependencies across different molecular layers that would otherwise be missed by analyzing individual data types. Nonetheless, the task of effectively managing and processing the vast amount of information present in multiomics data remains a challenge. One of the key challenges lies in selecting the essential variables or features that are most relevant to the underlying biological processes or phenomena of

interest. Variable selection becomes crucial to reduce noise, eliminate redundancy, and focus on the most informative features, thereby enhancing interpretability, computational efficiency, and the discovery of biomarkers or predictive models. This selection process challenge presents the first focus of this PhD.

5.6.2 Contribution

Description After reviewing the state-of-the-art, and mainly focusing on recent works [FLL+20, LFW+19], our aim is to develop an innovative feature selection method using multi-agent deep interactive reinforcement learning. The primary drawback of existing reinforcement-based approaches is their narrow focus on a single objective, primarily centered around accuracy improvement. However, when dealing with attribute selection, especially in the context of multiomics data, we encounter multiple conflicting objectives. For example, we need to balance accuracy maximization, minimizing the number of attributes, and maximizing feature relevance. Unlike current state-of-the-art methods, our proposed approach will address these diverse objectives in feature selection.

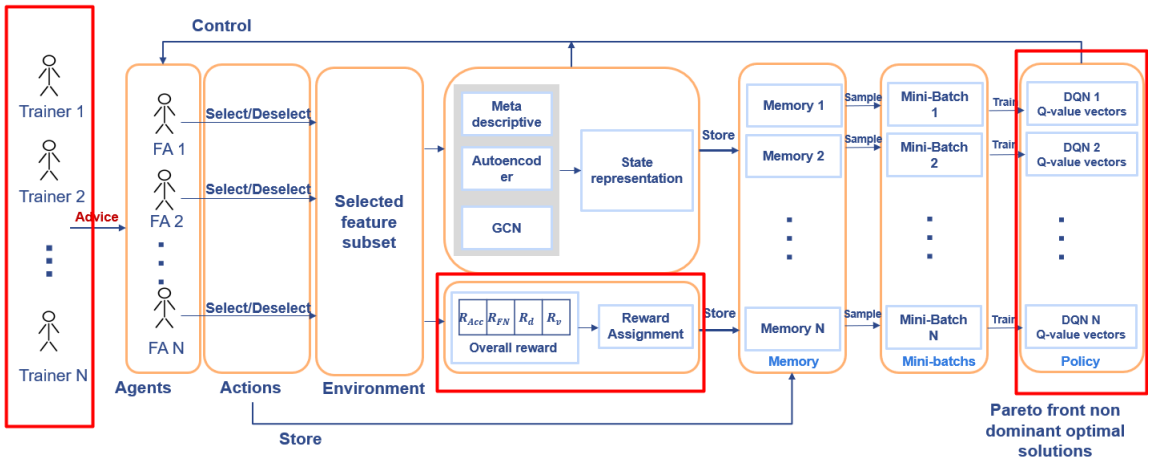

Figure 5.2: An overview of the multi-agent deep interactive reinforcement learning proposed solution

Additionally, our approach builds upon interactive deep reinforcement learning, which has demonstrated its effectiveness in accelerating exploration through the utilization of external knowledge. In this method, the agents' decisions are influenced by a trainer who provides appropriate guidance for each agent's actions. Existing approaches commonly employ a single trainer or maintain consistency throughout the learning process, resulting in a repetitive advice at each iteration. To overcome this limitation, our proposal suggests varying the trainers at each iteration to introduce diversity in decision-making. This strategy ensures that the same advice is not repeated at every learning step. By incorporating these enhancements, we strive to enhance the performance and effectiveness of feature selection in multiomics data analysis. A high level representation of our proposed solution is given in Figure 5.2.

Key conclusions The first technical contribution of this work is currently under development and is a work in progress.

5.7 Conclusion

The focus of this chapter was to outline the various contributions and work in progress in the field of knowledge discovery using biological computation and deep learning. These contributions encompass both practical applications and theoretical studies. The next chapter will delve into the third research direction of this Habilitation Thesis.

Chapter 6

Personalizing healthcare: Focus on real-world applications

6.1 Introduction

In this chapter, some of the research presented in the Habilitation Thesis is put into practice through a combination of secondments and collaborations with non-academic partners. The chapter presents two real-world applications: (i) cancer incidence prediction in epidemiology (Section 6.2), and (iii) personalized approaches to treating sepsis (Section 6.3). The conclusion is given in Section 6.4.

6.2 A case study in epidemiology

This section outlines the practical application that was carried out in the context of my MSC project, which was also integrated into the EPIDEMIUM project. Further details about the projects can be found in Chapter 2. The research findings presented in this case study were the product of a collaborative effort involving healthcare specialists from the Centre d'Épidémiologie Clinique, Hôpital Hôtel Dieu de Paris, France.

6.2.1 Context and motivation

Epidemiology is a sub-field of public health that looks to determine where and how often disease occur and why. It is more formally defined as the study of distributions (patterns) and determinants (causes) of health related states or events within a specified human population, and the application of this study to managing health problems [Woo13]. The ultimate goal of epidemiology is to apply this knowledge to the control of disease through prevention and treatment, resulting in the preservation of public health. In this context, epidemiologists study chronic diseases such as arthritis, cardiovascular disease such as heart attacks and stroke, cancer such as breast and colon cancer, diabetes, epilepsy and obesity problems. To conduct such studies, one of the most important considerations is the source and content of data, as this will often determine the quality of the results. As a general rule, the larger the data, the more accurate the results, since a larger sample is less likely to, by chance, generate

an estimate different from the truth in the full population. This leads epidemiologists to deal with big data which is however not a feasible task for them [MWES15].

Data analysis assists epidemiologists to investigate and describe the determinants and distribution of disease, disability, and other health outcomes and develop the means for prevention and control. Meanwhile, in epidemiology, feature reduction is a main point of interest and focusing on this phase is crucial as it often presents a source of potential information loss. In this case study, we mainly focused on the use of a feature selection technique, specifically a filter technique, instead of a feature extraction technique. This is crucial to preserve the semantics of the features in the context of cancer incidence prediction.

As previously discussed in Chapter 4 — Section 4.2.1, many feature selection techniques were proposed in literature [BKRA+16] and these have several limitations specifically when it comes to big data. To overcome their limitations, it was interesting to look for a filter method that does not require any external or supplementary information to function properly within the big data context. In this context, the application of the Sp-RST solution [DZBL17, DZBL20] was considered (discussed in Chapter 4 — Section 4.2.2) as a data mining technique within a case study in epidemiology and cancer incidence prediction.

6.2.2 Application

Description The OpenCancer[1] organization gathers people working on cancer prediction issues. It aim is to provide tools aimed at helping health authorities to take public policy decisions in terms of cancer prevention. OpenCancer has linked and merged data from the World Health Organization (WHO)[2], World Bank (WB)[3], the International Labour Organization (ILO)[4] and the Food and Agriculture Organization (FAO)[5] of the United Nations to build a large data set covering 38 countries and many regions within these countries between 1970 and 2002. For this application, OpenCancer provided a first version of the database restricted to the WHO, WB and FAO sources. For this application the single cancer type which has been considered is the colon cancer. To help epidemiologists in their decision making, we proceeded as follows:

- *Data cleaning*: The first version of this sub-database suffers from a vast number of missing cells due to the lack of information in the available repositories. To fix this issue prior to running any learning model, OpenCancer had discarded every feature exhibiting a missing data ratio higher than 50% and imputed other missing data with

[1] https://github.com/orgs/EpidemiumOpenCancer/

[2] http://www.who.int/en/

[3] http://www.worldbank.org/

[4] http://www.ilo.org/global/lang--en/index.htm

[5] http://www.fao.org/home/fr/

a standard mean strategy. The resulting database — merged from both FAO and WB, including the incidence provided from WHO — includes 3 365 features and 45 888 records. To measure the occurrence of the colon cancer disease in the population, the number of new cases of the disease within a population occurring over 1970 and 2002 is used, referring to the incidence measure.

- *Feature selection*: Once the database is ready for use, our developed Sp-RST distributed version [DZBL17, DZBL20] was applied to select the most important risk factors from the input database.

  Sp-RST was able to reduce the considered epidemiological database from 3 364 risk factors to only around 840 features.

- *Predictive modelling*: For colon cancer incidence prediction, the distributed version of the Random Forest Regression model[6] was used.

Key conclusions The main aim of our experiments was to demonstrate that Sp-RST is relevant to real-world applications as it can offer insights into the selected risk factors, speed up the learning process, ensure the performance of the colon cancer incidence prediction model without a significant information loss, and simplify the learned model for epidemiologists.

Based on our obtained results, reported in [DZS+18], 10 different reducts were generated and presented to the epidemiologists. We noticed that there is only a small variation in the distribution of the selected risk factors. We depicted the ratios of each category within the set of selected risk factors for each database (FAO & WB) as presented in Figure 6.1, and for an overall view as presented in Figure 6.2, the distribution of the categories of the combined data sets.

Based on these results, epidemiologists confirmed again that the selected risk factors are expected to appear in each of their corresponding databases (though potentially with a different overall distribution). This supports that Sp-RST can determine the key risk factors among a large set of features. Meanwhile, epidemiologists highlighted that a higher average or proportion does not necessarily mean that a risk factor is more important than another. Indeed, no firm conclusions on the influence of one factor on the colon incidence prediction can be drawn based on this information, only. Thus, from an epidemiological perspective, the risk factors selected by Sp-RST should be further coupled with other sources of data to complete the analysis and to be able to draw specific conclusions.

Adding to this, we investigated the selected risk factors per iteration (for all the 10 Sp-RST iterations). Each iteration of Sp-RST reflects a possible reduced set of risk factors on which the prediction of the colon cancer incidence can be made. We noticed that the risk factors

[6] https://spark.apache.org/docs/1.5.0/api/java/org/apache/spark/ml/regression/package-frame.html *{RandomForestRegressionModel, RandomForestRegressor}*

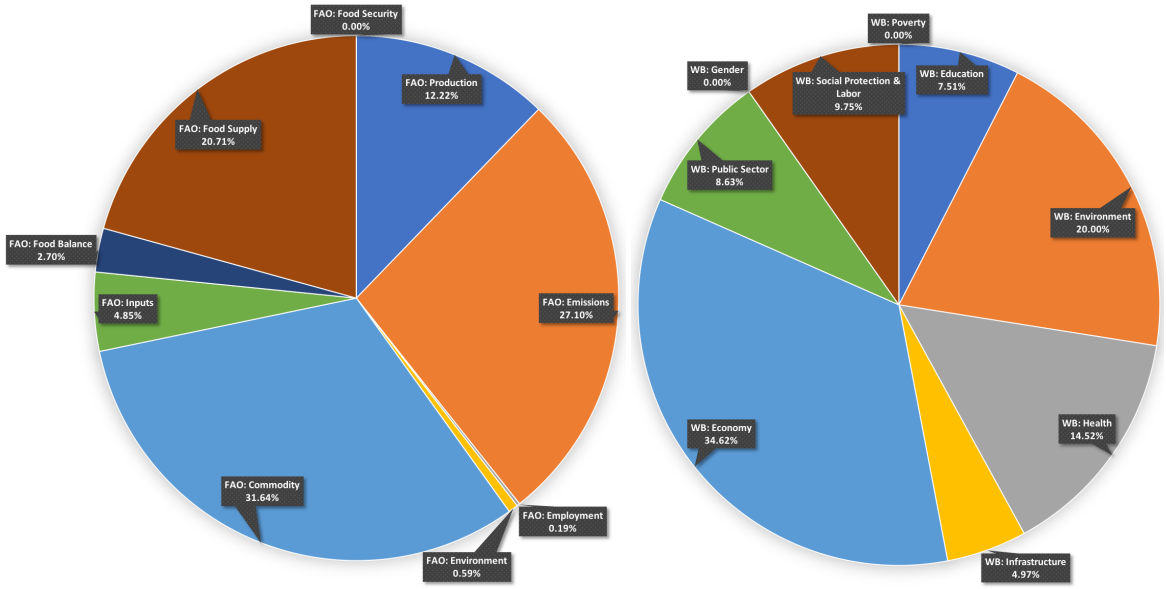

Figure 6.1: Distribution of the categories of risk factors selected by Sp-RST; split by data set: FAO (left) and World Bank (right)

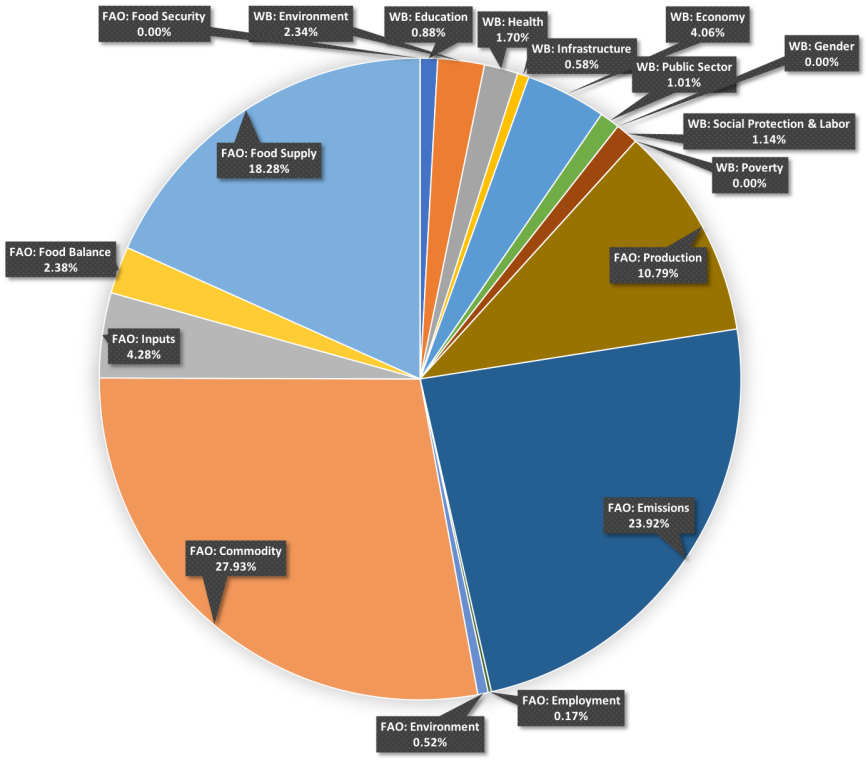

Figure 6.2: Distribution of the categories of the combined data set

partially overlap within the 10 iterations. An example of the 10 reducts for the FAO data set is presented in Figure 6.3. This is interesting from an epidemiological point of view as it can influence the consideration of other possible risk factors, which appear with different distributions. Indeed, the overlap between the selected risk factors from one iteration to another may call the attention of the epidemiologist in cases where a firm decision is taken with respect to a specific risk factor. These results are considered to be very important for the epidemiologists as they help them in the decision making process.

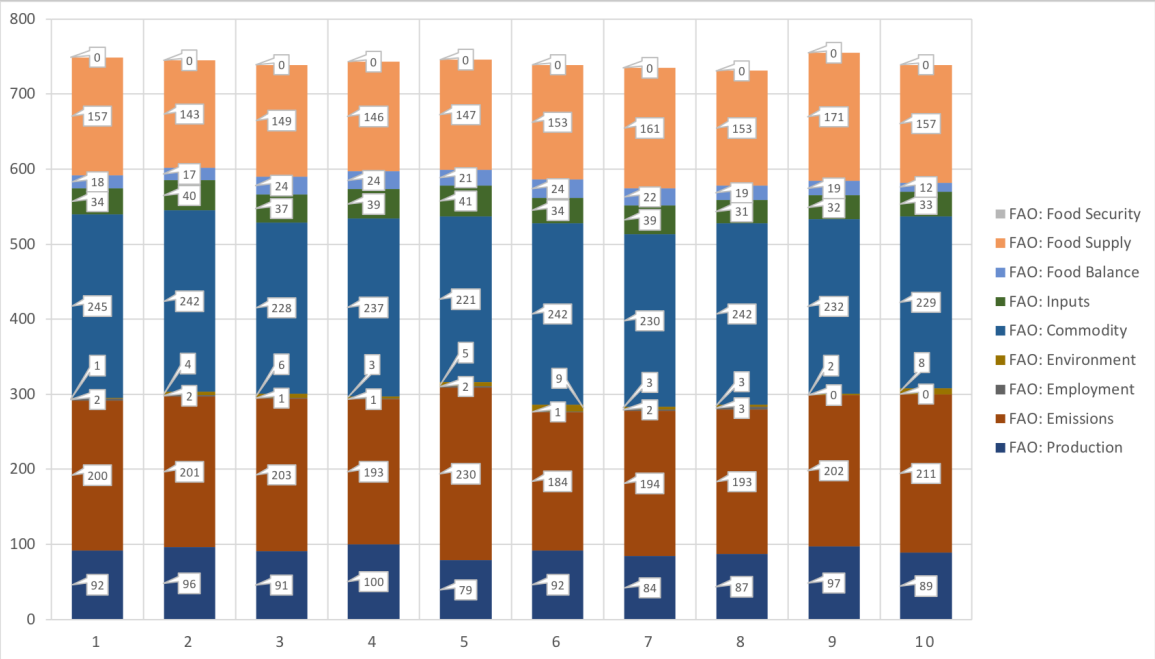

Figure 6.3: Categories of the selected risk factors in the FAO data set for each of the 10 iterations

With respect to our obtained results [DZS+18], we concluded that the quality of the obtained regression models is comparable (full data base vs reduced database). However, the reduced data set improves the execution time to determine the regression model considerably (by almost a factor of 5). Moreover, a data set with only around 840 risk factors is much easier to interpret and handle by epidemiologists. We therefore argue that the reduction process is appropriate in the considered context.

From our analyses, we concluded that using feature selection in the considered context is highly beneficial. The data set obtained is much easier to interpret and still yields comparable results.

6.2.3 Dissemination of results

The results of this case study were published at the *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2018* (ECML PKDD 2018)[DZS+18]. This application was also presented in some forums organized by

the European Commission such as the *Marie Skłodowska-Curie Actions Falling Walls Lab*, Brussels, and in other seminars such as the *Healthcare seminar, at Institut du Cerveau et de la Moelle épinière – ICM, Hôpital Pitié Salpêtrière*, Paris.

6.3 Personalized approaches to treating sepsis

In this section, a special focus is given to work that is being conducted in the frame of the SEPSIS and RECORDS projects (Chapter 2 – Table 2.4, Table 2.5). The research findings presented in this work resulted from a collaborative effort with healthcare specialists from the *Assistance Publique–Hôpitaux de Paris (APHP)*. Additionally, this work is part of my co-supervision of (i) a software engineer, and (ii) a postdoc, and my current co-supervision of (iiii) a PhD student (Chapter 3, Section 3.2.1).

6.3.1 Context and motivation

Sepsis is a life-threatening organ dysfunction caused by dysregulation of the host's response to infection that necessitates special medical treatment in the intensive care units of hospitals [SDS+16]. The sepsis disease life cycle involves three main stages: Systemic Inflammatory Response Syndrome (SIRS), severe sepsis, and septic shock [BBC+92]. In order to fight infection, the body releases additional immune system chemicals into the bloodstream when they are damaged. This is called the released immune system stage of sepsis, which can be extremely dangerous if it progresses rapidly. Severe sepsis occurs if the initial sepsis is not treated or does not meet treatment and may impact organ function. The symptoms of septic shock are similar to those of severe sepsis, but they also include a significant drop in blood pressure. This drop in blood pressure can lead to heart failure, stroke, other organ failures, respiratory failure, and even death [LHB+13]. The global sepsis case number is difficult to determine. An estimation has been made in 2017 indicating there were 48.9 million cases and 11 million deaths due to sepsis recorded worldwide, which represented about 20% of all deaths worldwide [RJA+20]. Also, maternal sepsis happens when sepsis occurs during pregnancy, during or after childbirth, or after a miscarriage. We also talk about neonatal sepsis when newborns are affected by the sepsis disease.

With the significant progress of the clinical best practices and the pharmaceutical industry, the risk of death was considerably reduced. However, the number of death cases still increases depending on the global number of varied hospitalized cases [NHR+18]. Therefore, the major challenge behind the sepsis mortality decrease is how to administrate the right treatment at the right time.

6.3.2 Applications

In this part, three initial contributions will be described. The first one presents a data mining methodology aiming at selecting the most efficient prediction model for predicting

Corticosteroid (CS) responsiveness in sepsis patients. The second contribution focuses on clustering sepsis data in the presence of missing values. The third contribution describes the context in which the newly multi-objective multi-agent deep interactive reinforcement learning approach (described in Chapter 5, Section 5.6) will be applied.

Application 1: Corticosteroid sensitivity detection in sepsis patients using a personalized data mining approach: a clinical investigation

Description Corticosteroid therapy has been shown to be related to the majority of sepsis patients in whom age, sex, disease severity, type of infection, source of infection, or type of pathogen do not influence survival benefit. However, the response to corticotherapy depends on the patient. The precise factors and biomarkers of responsiveness are various and complex [ABAB+19].

The purpose of this first contribution is to conduct a thorough clinical investigation on patients' responsiveness to CS by applying a data mining approach. Sepsis patients' data have been gathered from the APHP, with the goal to predict if sepsis patients' are CS sensitive or CS resistant; thus contributing to a better understanding of this condition. Collecting information on patient demographics, health outcomes, and samples, led to the construction of a first sepsis cohort, dubbed APPROCCHS [ARBB+18, ABC+16].

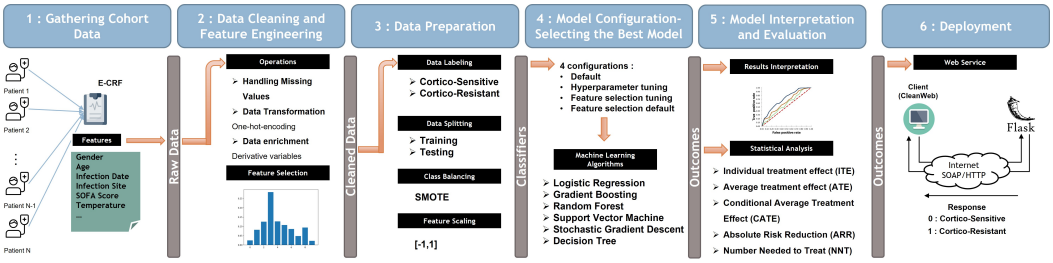

Figure 6.4: The applied KDD pipeline

The applied data mining pipeline, composed of six steps, is depicted in Figure 6.4. The first step in this process covers the acquisition of patients' sepsis related data. Next, data preprocessing is performed covering data cleaning and feature engineering. Features are selected for further processing based on whether or not the patient received corticosteroid treatment. Data preparation presents the third step. It includes labeling the data, splitting the data into training and testing, data scaling, and data balancing. In the fourth step, machine learning algorithms are applied to choose the best-performing one in detecting patient CS sensitivity. This step includes four configurations. Then, in order to interpret and evaluate the considered models, statistical analyses are performed using different measures. Finally, we deployed our selected machine learning model within the APHP information system. Our model was packaged within an Application Programming Interface (API) using Flask. It takes as input the patient's clinical data and returns whether the patient is sensitive or resistant to Corticosteroids.

From a clinical and data mining perspectives, we aim to answer the following questions:

- **RQ1:** Can predictive models recognize sepsis responders and non-responders to corticosteroid treatment?
- **RQ2:** How can patient features affect the accuracy of the obtained results?
- **RQ3:** What is the corticotherapy treatment effect at the individual patient level and at all treated patients?
- **RQ4:** Can the learned model (i.e., the signature) be generalized to different sepsis cohorts?

The APROCCHS cohort comprises data of 1240 patients, described using 5645 health characteristics (i.e., risk factors) with a specification if they were treated with corticosteroid or placebo. Since we need to know the resistance/sensitivity of the patient to the treatment before actually taking the drug, only data before the hospitalization (i.e., Day 0) have been considered to build the signature. However, it is worth noting that additional features until Day 2 of hospitalization have been also considered in our experimental setup, in order to provide a more comprehensive understanding of the patients' health status during treatment with corticotherapy, and to study the impact that such features (from Day 0 to Day 2) can have on the prediction generated by the considered machine learning models.

After enrolling a patient in the study on Day 0, he/she begin receiving corticosteroid treatment or a placebo every 4 to 6 hours while monitoring a series of features that indicate the patient's progress. Each patient is monitored for 90 days and feature values are recorded daily. The APHP medical experts have established clear criteria for determining whether a patient will respond to corticotherapy or not [ARBB+18, ABC+16]. Specifically, patients are classified as cortico-sensitive (i.e., responders) if all of the following four criteria are met after 14 days of treatment:

- The patient did not die.
- The vasopressor treatment is absent for at least 24 hours.
- The patient is free from mechanical ventilation for a least 24 hours.
- The SOFA score is less than 6.

If the criteria are not met, the treatment response is considered negative, meaning the patient is cortico-resistant or a non-responder. Therefore, data is labeled as 1 or 0, indicating whether the patient responded or did not respond to the treatment.

Key conclusions Two experiments were conducted to study the effectiveness of our approach to identify whether a sepsis patient is sensible or resistant to CS treatment. In the first experiment, and to answer **RQ1**, we carried out a thorough study of several machine learning models and compared their performance in terms of recognizing CS responsiveness.

Additional configurations have been included in order to further study the performance of our models. From the obtained results, we noticed that when the most relevant features were selected in combination with hyperparameter tuning, a machine learning model achieves the highest level of performance. Logistic Regression is considered to be the best model for our study with an AUC value of 72%; showing good performance in distinguishing sepsis responders and non-responders to CS treatment.

The second experiment is devoted to exploring the generalizability of the Logistic Regression model by testing it using additional data collected over time or from different patient groups belonging to a different cohort: RECORDS. RECORDS is an observational cohort characterized by the fact that most patients were affected by COVID-19; contrary to APPROCCHS. By analyzing the findings obtained from the model's variants, regarding only patients who got the CS treatment, and to answer **RQ2**, we found that when more information is added to the model (Day 0, Day 1, Day 2, and the difference between Day 2 and Day 1), the prediction results are improved. Our best-performing model achieved an AUC score of 74%. Moreover, to answer **RQ4** of whether the learned model can be generalized to different sepsis cohorts, we trained the model on APROCCHS and tested it on the RECORDS cohort. By interpreting the obtained results, we can conclude that the answer is negative. Thus, patient characteristics (such as those affected by COVID-19) can completely reduce the performance of a learned model; hence, the model cannot be generalized to different cohorts.

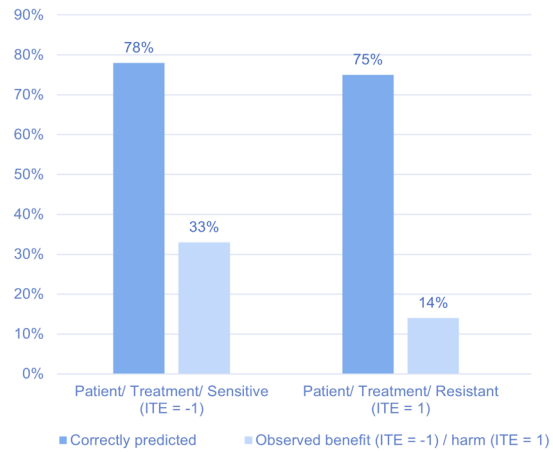

Figure 6.5: Percentage ITE with APROCCHS

Now to respond to **RQ3**, which focuses on the impact of CS on both individual treated patients and all treated patients as a whole, we evaluated the statistical obtained results. The reported findings are based on patients who were correctly predicted by the models, ensuring the accuracy and reliability of the results. The analysis of the results for APROCCHS for instance revealed that 33% of sepsis patients have shown an improvement (ITE = -1) while 14% of patients experienced a negative effect (ITE = 1) from the treatment; as can be seen in Figure 6.5. This information is important as it can assist medical professionals in providing

early interventions for resistant patients.

Application 2: Application of a game-theoretic rough sets three-way based approach for clustering sepsis data with missing values

Description Performing data mining tasks in the medical domain is a key challenge. This is mainly due to the uncertainty present in the patients' data such as the existence of missing data. In this contribution, we focused on this challenge.

Specifically, in the medical domain, the fact of imputing or deleting the tuples containing missing values may significantly influence the results. Pre-processing missing values may jeopardizes the quality and reliability of the machine learning results. Thus, a more appropriate and suitable strategy to handle missing values is to equip the machine learning model with an internal mechanism able to handle data with missing values. In this contribution, the issue and challenge of missing data in sepsis patients' data is tackled by the application of Game-Theoretic Rough Set (GTRS) as a three-way decision approach [AAY18].

As described in [AAY18], the players, the strategies, and the payoff or utility functions, are the three components which are needed to be defined to analyze problems with GTRS. The goal of this game is to enhance the clustering quality of data containing missing values. As defined in [AAYA18], this goal can be approached from the perspective of a trade-off between accuracy and generality of the clustering; where *Accuracy* refers to how well we cluster objects with missing values, whereas *generality* refers to the fraction of objects that were clustered in the first place. The game's ultimate objective and goal should be reflected in the players. In this regard, the players in this game present the clustering's accuracy and generality features. The strategies denote the different actions that a player can take in a game. To maximize her/his rewards/benefits, each player adopts a strategy. The outcomes of choosing a specific strategy are measured using a payoff function. The utility function is defined to reflect a player's potential performance gains or benefits from pursuing a specific strategy.

Table 6.1: Payoff table for the game

		Generality (G)		
		$s_1 = \alpha\downarrow$	$s_2 = \beta\uparrow$	$s_3 = \alpha\downarrow\beta\uparrow$
Accuracy (A)	$s_1 = \alpha\downarrow$	$u_A(s_1, s_1)$, $u_G(s_1, s_1)$	$u_A(s_1, s_2)$, $u_G(s_1, s_2)$	$u_A(s_1, s_3)$, $u_G(s_1, s_3)$
	$s_2 = \beta\uparrow$	$u_A(s_2, s_1)$, $u_G(s_2, s_1)$	$u_A(s_2, s_2)$, $u_G(s_2, s_2)$	$u_A(s_2, s_3)$, $u_G(s_2, s_3)$
	$s_3 = \alpha\downarrow\beta\uparrow$	$u_A(s_3, s_1)$, $u_G(s_3, s_1)$	$u_A(s_3, s_2)$, $u_G(s_3, s_2)$	$u_A(s_3, s_3)$, $u_G(s_3, s_3)$

Table 6.1 presents an example of strategies of the players. Each cell in Table 6.1 corresponds to a strategy profile, (s_m, s_n), where s_m represents player A's strategy and s_n represents player G's strategy. The goal of each player is to choose a strategy that configures the (α, β) thresholds in order to maximize her/his utility. $u_A(s_m, s_n)$ and $u_G(s_m, s_n)$ are the payoffs

for players A and G, respectively, according to the strategy profile (s_m, s_n).

The logic in the game is that a player chooses a strategy with a larger payoff over other strategies with a lower payoff. Therefore, the Nash equilibrium needs to be sought. As demonstrated in [AAYA18], when different thresholds are used in the game, the properties of accuracy and generality are influenced differently. Consequently, changes and variations in thresholds can be considered as feasible strategies. Once the optimal thresholds α and β have been determined, these are applied to cluster each data object into three classes: (i) inside the cluster, i.e., CS sensitive, (ii) outside the cluster, i.e., CS resistant, and (iii) Uncertain cluster where further investigation on the data object is needed. The three regions are defined as follows.

$$Inside(c_k) = \{o_i \in U | e(c_k, o_i) \geq \alpha\}, \tag{6.1}$$

$$Outside(c_k) = \{o_i \in U | e(c_k, o_i) \leq \beta\}, \tag{6.2}$$

$$Partial(c_k) = \{o_i \in U | \beta < e(c_k, o_i) < \alpha\}, \tag{6.3}$$

Key conclusions Our experimental procedure involved simulating missing values in the APPROCCHS cohort to examine how well the algorithm performs when more data is missing. We considered 10 risk factors (among the 24) as a first investigation. Based on the results, we observed that the trade-off between accuracy and generality is maintained when using the GTRS algorithm, even when the number of missing values increases. Furthermore, we noted that the output of the GTRS algorithm remains relatively stable regardless of the initial strategies used during the initialization phase.

After applying Equations 6.1, 6.2, and 6.3 to assign patients to clusters, using the selected optimal α and β parameters, we observed that the algorithm's non-deterministic nature resulted in some sepsis patients being found in the partial region. This means that these sepsis patients could not be clustered to a specific region; despite that we had their correct label in the cohort. One possible explanation to these preliminary results is that we have only considered 10 risk factors out of the 24 variables. Despite this, we still can consider that the initial results in terms of trade-off accuracy and generality are promising and indicate that GTRS has potential in addressing the issue of missing data in sepsis patients.

Application 3: Application of the multi-objective multi-agent deep interactive reinforcement learning technique for multiomics biomarker selection

In Chapter 6, Section, 5.6, an overview of the multi-objective multi-agent deep interactive reinforcement learning technique was given. This newly developed technique, dedicated for biomarker selection, will be applied to the SEPSIS/RECORDS (multi)-omics collected data; as presented in Figure 6.6. Currently, the availability of omics data in the APPROCCHS and RECORDS cohorts varies. It is important to note that not all patients have complete

omics information across all layers. However, we will begin our investigation by applying our technique to each available omic individually. This step will allow us to gain insights and assess the effectiveness of our approach within each specific omics domain.

By starting with the analysis of individual omics, we can uncover important biomarkers and patterns specific to each data layer. This focused exploration will provide valuable information about the molecular signatures associated with sepsis and help us understand the underlying mechanisms within each omics modality.

Once we have examined the available omics data individually, we will proceed with the integrative process. This involves combining the information from multiple omics layers to generate a more comprehensive view of the disease. Through the integration of diverse molecular data, we can identify cross-omics interactions, discover novel relationships, and unveil hidden connections that may not be evident when considering each omics layer in isolation.

The stepwise approach allows us to maximize the utility of the data we have at hand and progressively build a more holistic understanding of sepsis. It ensures that we make the most of the available information while laying the groundwork for a robust and comprehensive analysis of the disease using an integrative framework.

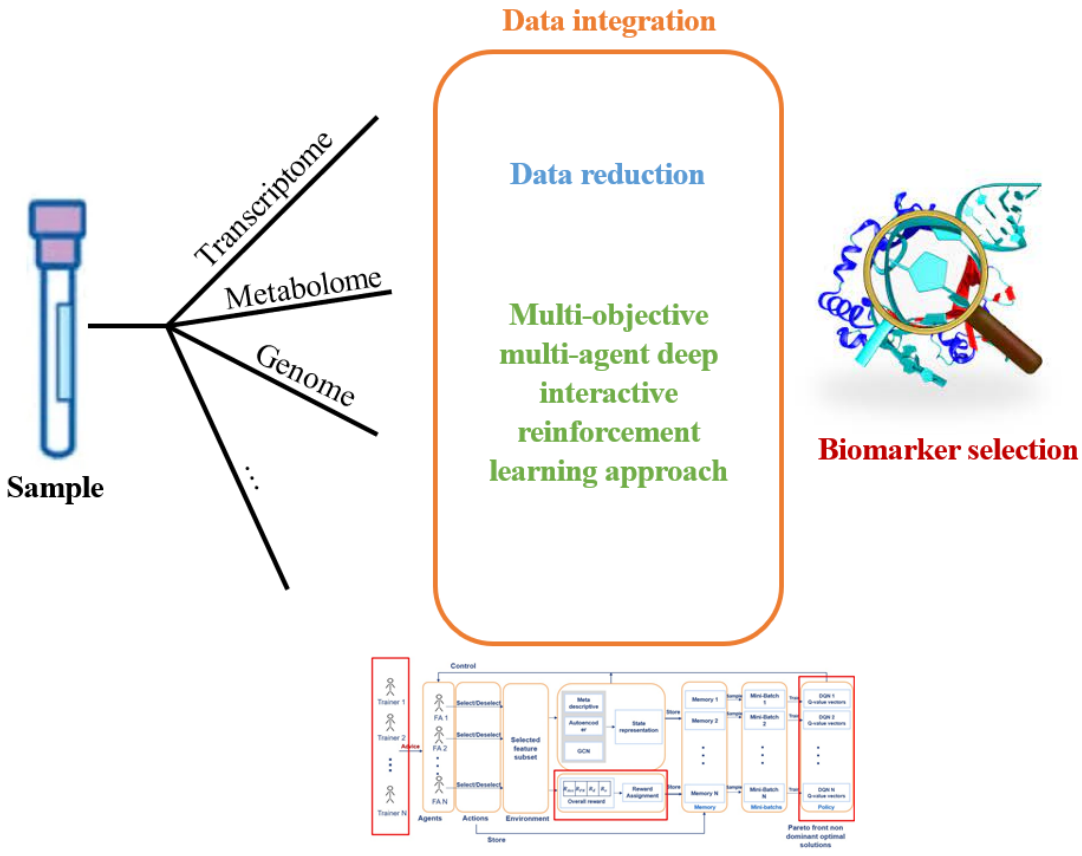

Figure 6.6: Approach of biomarker selection

6.3.3 Dissemination of results

The initial signature described in Section 6.3.2 presents the first release; deployed in APHP. It will be further updated and refined as new data and variables become available. The first contribution was submitted to the *Journal of Computer Methods And Programs In Biomedicine*. As for the second contribution described in Section 6.3.2, it was submitted to the *5th International Symposium on Rough Sets: Theory and Applications (RSTA'2023)* – part of the *18th Conference on Computer Science and Intelligence Systems (FedCSIS'2023)*. The third contribution is a work in progress.

6.4 Conclusion

In this chapter, the aim was to provide an overview of some real-world contributions in personalizing healthcare. The next chapter will present my list of publications.

Chapter 7

List of publications

Overall, my published works consist of **14 journal papers, 1 book, 3 book chapters, 31 peer-reviewed conference and workshop papers** (all of which were presented orally), and **9 seminar proceedings and press articles.**

7.1 Journal papers

1. Ben Amor, Omar, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. "Many-objective optimization of wireless sensor network deployment." Evolutionary Intelligence (2022): 1-17.
2. Zaineb Chelly Dagdia, and Ana Cristina Simões E Silva, "Effects of COVID-19 pandemic on education and society". STEM Education, 2,3,197,220, (2022)
3. Mkaouer, Mohamed Wiem, Tarek Gaber, and Zaineb Chelly Dagdia. "Effects of COVID-19 pandemic on computational intelligence and cybersecurity: survey." Applied Computing and Intelligence 2, no. 2 (2022): 173-194.
4. Jerbi, Manel, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. "Android malware detection as a bi-level problem." Computers & Security 121 (2022): 102825.
5. Zaineb Chelly Dagdia, Md Shamsuzzoha Bayzid, and Pavel Avdeyev, "Biological computation and computational biology: survey, challenges, and discussion". Artificial Intelligence Review, 54(6): 4169-4235 (2021).
6. Zaineb Chelly Dagdia, and Christine Zarges, "A detailed study of the distributed rough set based locality sensitive hashing feature selection technique". Fundamenta Informaticae 182(2): 111-179 (2021).
7. Zaineb Chelly Dagdia, Christine Zarges, Gael Beck, and Mustapha Lebbah, "A Scalable and Effective Rough Set Theory based Approach for Big Data Pre-processing". Knowledge and Information Systems: 62: 3321-3386 (2020).
8. Manel Jerbi, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said, "On the Use of Artificial Malicious Patterns for Android Malware Detection". Journal of Computers

& Security: 92: 101743 (2020).

9. Zaineb Chelly Dagdia, and Zied Elouedi, "A Hybrid Fuzzy Maintained Classification Method Based on Dendritic Cells". Journal of Classification: 37: 18-41 (2020).
10. Zaineb Chelly Dagdia, "A scalable and distributed dendritic cell algorithm for big data classification". Swarm and Evolutionary Computation Journal. 50: 1-13 (2019).
11. Zaineb Chelly Dagdia, "A Mobile Health Application Based on a Fuzzy QoS Web Service Component". Journal of Network and Innovative Computing: 7(1): 41-47 (2019).
12. Zeineb Chelly, and Zied Elouedi, "From the General to the Specific: Inducing a Novel Dendritic Cell Algorithm from a Detailed State-of-the-Art Review". International Journal of Pattern Recognition and Artificial Intelligence, IJPRAI 30(3): 1-31 (2016).
13. Zeineb Chelly, and Zied Elouedi, "A Survey of the Dendritic Cell Algorithm". Knowledge and Information Systems, KAIS 48(3): 505-535 (2016).
14. Zeineb Chelly, and Zied Elouedi, "Hybridization schemes of the fuzzy dendritic cell immune binary classifier based on different fuzzy clustering techniques". New Generation Computing, Springer 33(1): 1-31 (2015).

7.2 Book

1. Knowledge Discovery in Big Data from Astronomy and Earth Observation (AstroGeoInformatics), 1st Edition, eBook ISBN: 9780128191552, Elsevier. 2020 – *contribution:* Zaineb Chelly Dagdia, and Miroslav Mirchev, "When Evolutionary Computing meets Astro- and Geo- Informatics".

7.3 Book chapters

1. Juan Julian Merelo Guervos, Federico Liberatore, Antonio Fernandez-Ares, Ruben Hector Garcia-Ortega, Zeineb Chelly, Carlos Cotta, Nuria Rico, Antonio Mora Garcia, Pablo Garcia-Sanchez, Alberto Paolo Tonda, Paloma de las Cuevas, and Pedro A. Castillo, "The Uncertainty Quandary: A Study in the Context of the Evolutionary Optimization in Games and Other Uncertain Environments". Transactions on Computational Collective Intelligence, LNCS, TCCI 24: 40-60 (2016).
2. J.J. Merelo, Zeineb Chelly, Antonio Mora, Antonio Fernandez-Ares, Anna I. Esparcia-Alcazar, Carlos Cotta, P. de las Cuevas, and Nuria Rico, "A Statistical Approach to Dealing with Noisy Fitness in Evolutionary Algorithms". Studies in Computational Intelligence, Springer, pp 79-95 (2016)

3. Paloma de Las Cuevas Delgado, Zeineb Chelly, Antonio Mora Garcia, Juan Julian Merelo Guervos, and Anna Esparcia-Alcazar, “An Improved Decision System for URL Accesses Based on a Rough Feature Selection Technique”. Recent Advances in Computational Intelligence in Defense and Security, SCI, Springer, pp 139-167 (2016).

7.4 Conferences papers

1. Jerbi, Manel, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. “Immune-Based System to Enhance Malware Detection.” In IEEE 2023 Congress on Evolutionary Computation (to appear) (IEEE CEC 2023).

2. Zaineb Chelly Dagdia, and Vania Bogorny, “Towards a Granular Computing Framework for Multiple Aspect Trajectory Representation and Privacy Preservation: Research Challenges and Opportunities”, 17th Conference on Computer Science and Intelligence Systems (FedCSIS 2022).

3. Jerbi, Manel, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. “Malware Evolution and Detection Based on the Variable Precision Rough Set Model.” In 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 253-262. IEEE, 2022.

4. Manel Jerbi, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said, “Malware Detection Using Rough Set Based Evolutionary Optimization”, Proceedings of the 28th International Conference on Neural Information Processing, ICONIP’2021, Bali, Indonisia, Springer, pp. 634-641.

5. Zaineb Chelly Dagdia, Chiara Renso, Karine Zeitouni, and Nazim Agoulmine, “Towards a Federated Learning Approach for Privacy-aware Analysis of Semantically Enriched Mobility Data”, 1st workshop on Flexible Resource and Application Management on the Edge, FRAME@HPDC 2021, pp. 17-20 (online).

6. Paloma de las Cuevas, Pablo Garcia-Sanchez, Zaineb Chelly Dagdia, Maria Isabel Garcia-Arenas, and Juan Julian Merelo, “Automatic rule extraction using Genetic Programming in a “Bring Your Own Device” scenario”. The Leading European Event on Bio Inspired Computation, EVOSTAR’2020, Seville, Spain, pp. 54-69.

7. Zaineb Chelly Dagdia, Christine Zarges, Gael Beck, Hanene Azzag, and Mustapha Lebbah, “A Distributed Rough Set Theory Algorithm based on Locality Sensitive Hashing for an Efficient Big Data Pre-processing”. Proceedings of the IEEE Big Data Conference, Proceedings of the IEEE Big Data Conference, BigData’2018, Seattle, USA, IEEE, pp. 2597-2606.

8. Azam Hamidinekoo, Zaineb Chelly Dagdia, Zobia Suhail and Reyer Zwiggelaar, “Distributed Rough Set Based Feature Selection Approach to Analyse Deep and Hand-

crafted Features for Mammography Mass Classification", Proceedings of the IEEE Big Data Conference, BigData'2018, Seattle, USA, IEEE, pp. 2423-2432.

9. Zaineb Chelly Dagdia, Christine Zarges, Benjamin Schannes, Martin Micalef, Lino Galiana, Benoit Rolland, Olivier de Fresnoye, and Mehdi Benchoufi, "Rough Set Theory as a Data Mining Technique: A Case Study in Epidemiology and Cancer Incidence Prediction", European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML'2018, Dublin, Ireland, Springer, pp. 440-455.

10. Zaineb Chelly Dagdia, "A Distributed Dendritic Cell Algorithm for Big Data", Proceedings of the Genetic and Evolutionary Computation Conference, GECCO'2018, Kyoto, Japan, ACM, pp. 103-104.

11. Zaineb Chelly Dagdia, Christine Zarges, Gael Beck and Mustapha Lebbah, "Modèle de Sélection de Caractéristiques pour les Données Massives", Proceedings of the 15ème édition de l'atelier Fouille de Données Complexes, FDC'2018, Paris, France, pp 1-12.

12. Zaineb Chelly Dagdia, Christine Zarges, Gael Beck and Mustapha Lebbah, "Nouveau Modèle de Sélection de Caractéristiques basé sur la Théorie des Ensembles Approximatifs pour les Données Massives", Proceedings of the 18ème édition de la conférence internationale francophone Extraction et Gestion de Connaissances, EGC'2018, Paris, France, pp 377-378.

13. Zaineb Chelly Dagdia, Christine Zarges, Gael Beck and Mustapha Lebbah, "A Distributed Rough Set Theory based Algorithm for an Efficient Big Data Pre-processing under the Spark Framework". Proceedings of the IEEE Big Data Conference, BigData'2017, Boston, USA, IEEE, pp 911-916.

14. Zeineb Chelly, "Data Pre-processing Based on Rough Sets and the Link to Other Theories". Proceedings of the Second International Afro-European Conference for Industrial Advancement, AECIA'2015, Villejuif, France, Springer, pp 5-7.

15. Kaouther Ben Ali, Zeineb Chelly, and Zied Elouedi, "A New Version of the Dendritic Cell Immune Algorithm based on the K-Nearest Neighbors". Proceedings of the 22th International Conference on Neural Information Processing, ICONIP'2015, Istanbul, Turkey, Springer, pp 688-695.

16. Juan J. Merelo, Federico Liberatore, Antonio Fernandez-Ares, Ruben Hector Garcia-Ortega, Zeineb Chelly, Carlos Cotta, Nuria Rico, Antonio Miguel Mora, and Pablo Garcia-Sanchez, "There is Noisy Lunch: A Study of Noise in Evolutionary Optimization Problems". Proceedings of Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI 2015) - Volume 1: ECTA, Lisbon, Portugal, pp 261-268.

17. Zeineb Chelly, and Zied Elouedi, “A Study of the Data Pre-Processing Module of the Dendritic Cell Evolutionary Algorithm”. Proceedings of the 2nd International Conference on Control, Decision and Information Technologies, Codit’2014, Metz, France, pp. 634-639, IEEE.

18. Zeineb Chelly, and Zied Elouedi, “A two-leveled hybrid dendritic cell algorithm under imprecise reasoning”. Proceedings of the Genetic and Evolutionary Computation Conference, GECCO’2014, Vancouver, Canada, pp. 97-104, ACM.

19. Zeineb Chelly, and Zied Elouedi, “A Rough Information Extraction Technique for the Dendritic Cell Algorithm within Imprecise Circumstances”. Proceedings of the 8th Hellenic Conference on Artificial Intelligence, SETN’2014, Ioannina, Greece, pp. 43-56, Springer.

20. Zeineb Chelly, and Zied Elouedi, “Improving the dendritic cell algorithm performance using fuzzy-rough set theory as a pattern discovery technique”. Proceedings of the 5th International Conference on Innovations in Bio-Inspired Computing and Applications, IBICA’2014, Ostrava, Czech Republic, pp. 23-32, Springer.

21. Mohamed Amine Ben Yahmed, Mohamed Amine Bounenni, Zeineb Chelly, and Amir Jlassi, “New Mobile Health Application for an Ubiquitous Information System”, Proceedings of the 6th Joint IFIP Wireless and Information System, WMNC’2013, Dubai, United Arab Emirates, pp. 1- 4, IEEE.

22. Zeineb Chelly, and Zied Elouedi, “A New Hybrid Fuzzy-Rough Dendritic Cell Immune Classifier”, Proceedings of the 4th International Conference on Swarm Intelligence, IC-SI’2013, Harbin, China, pp. 514-521, Springer.

23. Zeineb Chelly, and Zied Elouedi, “A New Data Pre-processing Approach for the Dendritic Cell Algorithm Based on Fuzzy Rough Set Theory”, Proceedings of the Genetic and Evolutionary Computation Conference, GECCO’2013, Amsterdam, Netherlands, pp. 163 – 164, ACM.

24. Zeineb Chelly, and Zied Elouedi, “A Fuzzy-Rough Data Pre-processing Approach for the Dendritic Cell Classifier”, Proceedings of the 12th European Conference on Symbolic and Qualitative Approaches to Reasoning with Uncertainty, ECSQARU’2013, Utrecht, Netherlands, pp. 109 – 120, Springer.

25. Zeineb Chelly, and Zied Elouedi, “QR-DCA: A New Rough Data Pre-processing Approach for the Dendritic Cell Algorithm”, Proceedings of the 11th International Conference on Adaptive and Natural Computing Algorithms, ICANNGA’2013, Lausanne, Switzerland, pp. 140-150, Springer.

26. Zeineb Chelly, and Zied Elouedi, “Supporting Fuzzy-Rough Sets in the Dendritic Cell Algorithm Data Pre-processing Phase”, Proceedings of the 20th International

Conference on Neural Information Processing, ICONIP'2013, Daegu, South Korea, pp. 164-171, Springer.

27. Zeineb Chelly, and Zied Elouedi, "Further Exploration of the Hybrid Fuzzy-Rough Dendritic Cell Immune Classifier Phase", Proceedings of the 4th, IEEE International Conference on E-Health and Bioengineering, EHB'2013, Iasi, Romania, pp. 1- 4, IEEE.

28. Zeineb Chelly, Abir Smiti, and Zied Elouedi, "COID-FDCM: The Fuzzy Maintained Dendritic Cell Classification Method", Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing, ICAISC'2012, Zakopane, Poland, pp. 233-241, Springer.

29. Zeineb Chelly, and Zied Elouedi, "RC-DCA: A New Feature Selection and Signal Categorization Technique for the Dendritic Cell Algorithm Based on Rough Set Theory", Proceedings of the 11th International Conference of Artificial Immune Systems, ICARIS'2012, Taormina, Sicily, Italy, pp. 152-165, Springer.

30. Zeineb Chelly, and Zied Elouedi, "RST-DCA: A Dendritic Cell Algorithm Based on Rough Set Theory", Proceedings of the 19th International Conference on Neural Information Processing, ICONIP'2012, Doha, Qatar, pp. 480-487, Springer.

31. Zeineb Chelly, and Zied Elouedi, "Further Exploration of the Fuzzy Dendritic Cell Method", Proceedings of the 10th International Conference of Artificial Immune Systems, ICARIS'2011, Cambridge, England, UK, pp. 419-432, Springer.

32. Zeineb Chelly, and Zied Elouedi, "FDCM: A Fuzzy Dendritic Cell Method", Proceedings of the 9th International Conference of Artificial Immune Systems, ICARIS'2010, Edinburgh, Scotland, UK, pp. 102-115, Springer.

7.5 Seminars proceedings

1. Zaineb Chelly Dagdia, "Optimized Framework based on Rough Set Theory for Big Data Preprocessing in Certain and Imprecise Contexts", Proceedings of The 5th MCAA Annual Conference and General Assembly 2018, Leuven, Belgium[1].

2. Zaineb Chelly Dagdia, "Optimized Framework based on Rough Set Theory for Big Data Pre-processing in Certain and Imprecise Contexts" – Marie Sklodowska-Curie Project: Open Problems'. Dagstuhl Seminar 17381, Recent Trends in Knowledge Compilation 2017, Dagstuhl, Germany, pp.70[2].

[1] https://www.mariecuriealumni.eu/sites/default/files/mcaa_book_of_abstracts18_v8.pdf?utm_source=facebook&utm_campaign=smis&utm_content=boa

[2] http://dx.doi.org/10.4230/DagRep.7.9.62

7.6 Press articles

1. Debate "Education, Research & Innovation: developing concrete synergies" (2018)[3]
2. Marie Skłodowska-Curie Poem (2018)
3. Role Models for Mobility of Women Scientists (2018)[4]
4. Experience, Learn and Share at the Heidelberg Laureate Forum - Heidelberg Laureate Forum Media (2017)[5]
5. Experience, Learn and Share at the Heidelberg Laureate Forum - ACM's Women in Computing , ACM-W, Europe (2017)[6]
6. Mentioned in the ACM-W Connections. (2017)[7]
7. ACM-W Supporting, Celebrating and advocating for Women in Computing. (2014)[8]

[3] https://acmweurope.acm.org/debate-education-research-innovation-developing-concrete-synergies/

[4] http://www.therolemodels.net/wp-content/uploads/2018/06/Ebook-2018-5.pdf

[5] https://scilogs.spektrum.de/hlf/experience-learn-share-heidelberg-laureate-forum/

[6] https://acmweurope.acm.org/experience-learn-and-share-at-the-heidelberg-laureate-forum/

[7] https://women.acm.org/ACM-W-Connections-2017-01/

[8] https://women.acm.org/scholars/acm-w-scholars/zeineb-chelly/

Conclusion

Throughout this Habilitation Thesis, I have presented my research and career development. First, in Chapter 1, I gave a retrospective view of my professional and academic journey. In Chapter 2, I discussed my research aptitude, highlighting my experience in various scientific endeavors and my ability to engage in research and technology transfer activities. In Chapter 3, my aptitude of supervision was emphasized. Chapter 4 of my Habilitation Thesis details the contributions made in the first research direction, which focuses on utilizing granular computation and federated learning for knowledge discovery and privacy preservation. The second research direction is the subject of Chapter 5, which discusses my work in the field of artificial immune systems and evolutionary computation. The third research direction, discussed in Chapter 6, focuses on putting into practice some of the research conducted in this Habilitation Thesis. Finally, a list of all my publications was presented in Chapter 7. My research involved a collaborative endeavor with numerous academic and non-academic partners, along with the participation of PhD students and postgraduates. This research was carried out within several funded research projects.

As highlighted in this Habilitation Thesis, my research covers work in progress, and I am continuously exploring new avenues for advancing the research directions presented in Chapter 4, Chapter 5, and Chapter 6.

In my future research, I intend to further investigate privacy preservation for semantically enriched mobility data. The dynamic nature of mobility data, characterized by people's movement and evolving transportation patterns, poses a challenge for privacy-preserving techniques. It is crucial to develop adaptive methods that effectively protect privacy in these changing environments while ensuring robustness and accuracy. To address this challenge, I plan to explore the utilization of federated learning and granular computation techniques. Additionally, I aim to examine how these techniques can overcome specific challenges related to privacy preservation, including data aggregation, data heterogeneity, and data bias. By investigating the potential of federated learning and granular computation, I hope to contribute to the development of effective solutions that address privacy concerns in semantically enriched mobility data.

I am also highly committed to developing innovative machine/deep learning techniques that specifically address the challenges associated with medical/health data. This encompasses the creation of cutting-edge multi-modal machine/deep learning methods, where my primary objective is to leverage the potential of diverse data modalities in order to enhance model accuracy. My focus lies in investigating novel techniques for effectively integrating and

combining data from various sources, such as images and text, to develop models that can better capture and represent complex phenomena in the medical field. In addition, I have a strong emphasis on developing models that are explainable, meaning they can provide insights and justifications for their predictions or decisions. This is crucial in the healthcare domain, where interpretability and transparency are paramount for building trust and understanding. Furthermore, I aim to explore the realm of Zero/Few-shot learning, which involves training models to make accurate predictions when presented with new or scarce data instances. This direction will enable the models to generalize well even with limited labeled examples, which is particularly valuable in medical scenarios where obtaining large annotated datasets can be challenging. Each of these envisioned future directions brings forth its own set of hidden challenges and problems to be solved.

Part III

Appendices

Chapter 8

The dendritic cell algorithm

8.1 Introduction

As previously stated, the most prominent players of the Danger Theory are the Dendritic Cells (DCs). An inspiration from the DCs behavior led to the development of an immune algorithm termed the Dendritic Cell Algorithm (DCA) [GA05]. In this Chapter, the DCA is described first via its biological principals (Section 8.2), then via its algorithmic steps (Section 8.3), and finally via a worked example (Section 8.4).

8.2 Biological background

A DC is a type of antigen-presenting cell. DCs are in charge of catching, processing and revealing antigens to T-cells. They, also, express receptors on their surfaces to receive signals from their neighborhood. The behavior of DCs depends on the concentration of the signals received. As a result, they differentiate into three different maturity levels described as follows [LS02]:

- **Immature DCs:** On arrival in the tissue, DCs are found in an immature state. Here, immature DCs (iDCs) collect antigen which could be a “safe" molecule or something foreign. Furthermore, DCs can collect and sense the various signals that may be present in the tissue. Receipt of signals causes changes to the function, morphology and behavior of the iDC. In other words, the relative proportions and potency of the different signals lead to a full or partial maturation state of iDCs.
- **Mature DCs:** For an iDC to become a “mature DC" (mDC), the iDC must be exposed to a greater quantity of either PAMPs or DSs than SSs. Sufficient exposure to PAMPs and DSs causes the DC to cease antigen collection and migrate from the tissue to the lymph node. Most importantly, mDCs produce an inflammatory cytokine called “interleukin-12" which stimulates T-cell activation in order to be reactive to antigen presentation. Additionally, mDCs produce costimulatory molecules (CSMs) which are known to facilitate the antigen presenting process.
- **Semi-mature DCs:** In the presence of apoptosis conditions, exposure to SSs diverts

the iDC to become a “semi-mature DC" (smDC). smDCs appear morphologically very similar to mDCs and can present antigen, yet they do not have the ability to activate T-cells. The smDC produces “interleukin-10" which suppresses T-cells matching the presented antigen. Antigens collected with SSs are presented in a tolerogenic context and lead to unresponsiveness to those antigens.

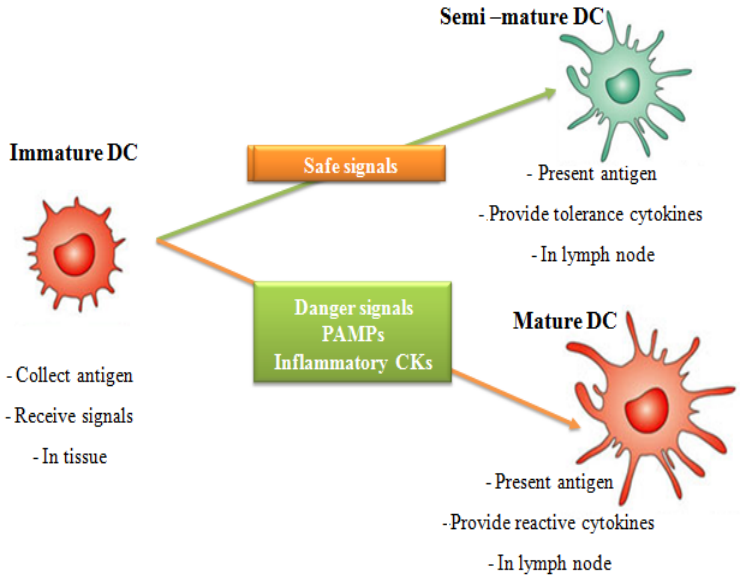

Figure 8.1: DCs behavior

Swapping from one DC state to another is dependent upon the receipt of different signals throughout the initial state (iDC). This is shown in Figure 8.1. The migration to the mature state or to the semi-mature state depends on the concentration of the input signals received from the environment. Immunologically, if the concentration of SSs is greater than the other categories of signals then DCs migrate to the semi-mature state. However, if the concentration of PAMPs and DSs is greater than the concentration of SSs then DCs migrate to the mature context.

8.3 Algorithmic details

To transform the abstract model of DC biology into an immune-inspired algorithm, it must be formalized into the structure of a generic algorithm and into a set of logical processes. A generic form of the DCA binary classifier is described by Algorithm 1.

For the ease of analysis, the algorithm is divided into the following four main phases:

1. *Pre-Processing & Initialization phase* - Line 1 to Line 5;
2. *Detection phase* - Line 6 to Line 13;
3. *Context Assessment phase* - Line 14 to Line 19;

Algorithm 1 The Dendritic Cell Algorithm

```
1: input: signals from all categories and antigens;
2: output: antigens plus context values;
3: for each DC do                          /* Pre-processing & Initialization phase */
4:    initialize DC;
5: end for
6: while CSM output signal < migration threshold do        /* Detection phase */
7:    get antigens;
8:    store antigens;
9:    get signals;
10:   calculate interim output signals;
11:   update cumulative output signals;
12: end while
13: cell location update to lymph node;
14: if semi-mature output > mature output then          /*Context Assessment phase */
15:   cell context is assigned as 0;
16: else
17:   cell context is assigned as 1;
18: end if
19: print collected antigens plus cell contexts;
20: for all antigens in total list do                        /* Classification phase */
21:   increment antigen count for this antigen type;
22:   if antigen context equals 1 then
23:       increment antigen type mature count;
24:   end if
25: end for
26: for all antigen types do
27:    MCAV of antigen type = mature count / antigen count;
28: end for
```

4. *Classification phase* - Line 20 to Line 28;

Pre-Processing and initialization phase The application of the DCA often requires a data pre-processing phase to appropriately map a given problem domain to the input space of the algorithm. The pre-processing stage involves two main steps: feature reduction and signal categorization. The process of feature reduction and signal categorization involves selecting/extracting the most interesting features from the original feature set of a given problem, and then categorizing each of these derived features into one of the defined signal categories of the DCA; i.e., PAMP, DS or SS. Some DCA works deal with involving users or experts in order to select or extract the most interesting features and map them into the appropriate signal categories. Other DCA works apply, principally, the Principal Component Analysis (PCA) statistical method for an automated data pre-processing task. More precisely, for feature reduction, DCA applies the PCA that selects the first “principal components" which reveal the internal structure of the given data with the focus on data variance. Once features are selected, PCA is applied to assign each attribute to its specific signal type.

For signal categorization, some DCA versions use one attribute among the selected set

of features and assign it to both PAMP and SS as they are both considered as positive indicators of an anomalous and normal signal. Using one attribute for these two signals requires a threshold level to be set: values greater than this can be classed as SS otherwise as PAMP. As for the DS attribute assignment and since the DS is less than certain to be anomalous, the combination of the rest of the selected attributes are chosen to represent it. Other DCA works handle the signal categorization step using the PCA ranking procedure. This is performed by the use of the PCA attributes' ranking in terms of variability. Once ranking is performed, the attributes are mapped into the DCA input signal categories by correlating the PCA ranking with the ranking of signal categories which is in the order Safe, PAMP, and Danger [GOGA09].

Detection phase Once the DCA data pre-processing stage is achieved, the algorithm calculates the SS, PAMP and DS values [Gre07] inducing a signal database and adheres these signals and antigen to fix the context of each DC. This is preformed during the detection phase. In fact, the input signals of the system which are pre-categorized as "PAMP", "danger" and "safe" are processed by the algorithm in order to get three output signals: costimulation signal (CSM), semi mature signal (smDC) and mature signal (mDC). The detection phase occurs within DCs of the immature state. The DC has the following three functions which are performed each time a single DC is updated:

1. *Sample antigen:* the DC collects signals and antigens from an external source (in this case, from the tissue) and places the antigen in its own antigen storage data structure (Line 7 to Line 9).

2. *Update input signals:* the DC collects values of all input signals present in the signal storage area.

3. *Calculate interim output signals:* at each iteration, each DC calculates three temporary output signal values from the received input signals. These signals are used to assess the state of the DC upon termination of the detection phase of a DC's life span. The three output signals of a DC perform two roles, to determine if an antigen type is anomalous and to limit the time spent sampling data.

To calculate the interim output signals, DCA applies the following weighted sum equation:

$$C_{[CSM,smDC,mDC]} = \frac{((W_{PAMP} * \sum_i PAMP_i) + (W_{SS} * \sum_i SS_i) + (W_{DS} * \sum_i DS_i))}{(W_{PAMP} + W_{SS} + W_{DS})} * \frac{1+I}{2} \tag{8.1}$$

Assuming that there are multiple signals per category, $PAMP_i$, DS_i and SS_i are the input signal values of category PAMP, danger and safe for all signals (i) of that category. W_{PAMP}, W_{SS} and W_{DS} represent the weights used for PAMP, SS and DS, respectively. I represents the inflammation signal. This equation is repeated three times, once per output signal. This is to calculate the interim output signal values for the CSM output, the smDC output and the mDC output. These values are cumulatively summed over time [Gre07]. The

weights used by the DCA are either derived empirically from the data or are user defined values.

Each DC in the population is assigned a migration threshold value upon its creation. Following the update of the cumulative output signals, a DC compares the value it contains for CSM with the value it is assigned as its migration threshold. If the value of CSM exceeds the value of the migration threshold then the DC is removed from the sampling area and its life span is terminated.

Context assessment phase Once the cell has migrated and through the context assessment phase, each DC has the ability to process and collect signals and antigens. Through the generation of cumulative output signals, the DC forms a cell context that is used to perform anomaly detection in the assessment of antigens. In fact, upon migration, the cumulative output signals are assessed and the greater of semi-mature or mature output signal becomes the cell context. This cell context is used to label all antigens collected by the DC with the derived context value of 1 or 0. This information is ultimately used in the generation of an anomaly coefficient which will be deal with in the final step; i.e., the classification phase.

Classification phase The derived value for the cell context is used to derive the nature of the response by measuring the number of DCs that are fully mature and is represented by the Mature Context Antigen Value (MCAV). The MCAV is used to assess the degree of anomaly of a given antigen. The closer the MCAV is to 1, the greater the probability that the antigen is anomalous. By applying thresholds at various levels, analysis can be performed to assess the anomaly detection capabilities of the algorithm. Those antigens whose MCAVs are greater than the anomalous threshold are classified into the anomalous category while the others are classified into the normal one.

8.4 DCA: An example

This example consists of applying DCA to the problem of credit management. An extract of the data set for this example is presented in Table 8.1.

Table 8.1: Bank database

Client	Age	Income	Number of credit cards	Duration of the loan	Credit
Client1	24	650	1	30	no
Client2	30	1000	3	10	no
Client3	36	1300	3	8	yes
Client4	20	600	1	20	no
Client5	32	900	2	13	yes

The dendritic cell algorithm selects, first of all, some attributes and pre-categorizes them as PAMP, DS and SS. Then, the obtained data set is transformed into a signal data set. An extracted set of the signal database is illustrated in Table 8.2.

Table 8.2: Signal data set

Client (antigen)	PAMP	SS	DS
Client1	100	100	0
Client2	0	0	100
Client3	20	50	40

To show the calculations under different input signal conditions, three iterations (cycles) with three sets of signals are shown. The derived output signal values are used to demonstrate how to perform the MCAV calculation for three different antigen types (Ag1, Ag2 and Ag3). In this example, three DCs are required, one for each iteration, termed DC1, DC2 and DC3 for the purpose of identification. Each DC is assigned an identical migration threshold value (tm) which is set to 100.

The sets of signals used in this example are presented in Table 8.2. The signal processing equation is the following:

$$C_{[CSM,smDC,mDC]} = (W_{PAMP} * PAMP) + (W_{SS} * SS) + (W_{DS} * DS)$$

The weights are presented in Table 8.3.

Table 8.3: Example of weights used for signal processing

	PAMP	SS	DS
CSM	2	1	2
smDC	0	0	1
mDC	2	1	- 1.5

The worked example is performed in the following itemized list:

1. We assume that the antigen vector (A) is the following:
   A = {Ag1; Ag1; Ag1; Ag1; Ag1; Ag2; Ag2; Ag2; Ag2; Ag3; Ag3; Ag3}

2. Cycle l = 0:
   DC samples antigens, so DC1 a(m) = {Ag1; Ag1; Ag1; Ag2; Ag2}
   DC samples input signals, so DC1 s(m) = {100; 100; 0}
   DC calculates output signals, so DC1 outputs:
   C_{CSM} = (100 * 2) + (100 * 1) + (0 * 2) = 300
   C_{smDC} = (100 * 0) + (100 * 0) + (0 * 1) = 0
   C_{mDC} = (100 * 2) + (100 * 1) + (0 * -1.5) = 300

For DC1, t(m) = 100, therefore this DC has now exceeded its migration threshold as the value for C_{CSM} is greater than t(m). Also, $C_{smDC} < C_{mDC}$ and therefore DC1 is assigned a cell context value of 1 indicating that its collected antigens may be anomalous.

3. By removing the antigens used by DC1, the antigen vector now consists of: A = {Ag1; Ag1; Ag2; Ag2; Ag3; Ag3; Ag3}

4. Cycle l = 1:
   DC samples randomly antigens, so DC2 a(m) = {Ag2; Ag2; Ag1}
   DC samples input signals, so DC2 s(m) = {0; 0; 100}
   DC calculates output signals, so DC2 outputs:
   C_{CSM} = (0 * 2) + (0 * 1) + (100 * 2) = 200
   C_{smDC} = (0 * 0) + (0 * 0) + (100 * 1) = 100
   C_{mDC} = (0 * 2) + (0 * 1) + (100 * -1.5) = -150
   For DC2, t(m) = 100, therefore this DC has now exceeded its migration threshold as the value for C_{CSM} is greater than t(m). Also, $C_{smDC} > C_{mDC}$ and therefore DC2 is assigned a cell context value of 0 indicating that its collected antigens are likely to be normal.

5. The antigen vector now consists of: A = {Ag1; Ag3; Ag3; Ag3}

6. Cycle l = 2:
   DC samples antigens, so DC3 a(m) = {Ag1; Ag3; Ag3; Ag3}
   DC samples input signals, so DC3 s(m) = {20; 50; 40}
   DC calculates output signals, so DC3 outputs:
   C_{CSM} = (20 * 2) + (50 * 1) + (40 * 2) = 170
   C_{smDC} = (20 * 0) + (50 * 0) + (40 * 1) = 40
   C_{mDC} = (20 * 2) + (50 * 1) + (40 * -1.5) = 30
   For DC3, t(m) = 100, therefore this DC has now exceeded its migration threshold as the value for C_{CSM} is greater than t(m). Indeed, $C_{smDC} > C_{mDC}$ and therefore DC3 is assigned a cell context value of 0.

7. Now antigens can be analyzed and the derived MCAV coefficients are shown in Table 8.4.

Table 8.4: Worked example of MCAV output

Antigen Type	num presentations	num mature presentations	MCAV
Ag1	5	3	0.6
Ag2	4	2	0.5
Ag3	3	0	0.0

8. To perform anomaly detection, a threshold must be applied to the MCAVs. This

threshold is a user defined parameter which requires some expert knowledge to define and is specific to the application. In this case, the anomaly threshold is defined by the bank manager and is set to 0.47. Therefore, client1 (Ag1) and client2 (Ag2) are classed as anomalous (they are not allowed to have a credit). However, client3 (Ag3) is classified as normal.

8.5 Conclusion

In this Appendix, the basic notions of the DCA were presented. First, its biological background was presented, followed by the DCA different algorithmic steps. Finally, a worked example of the algorithm was given.

Chapter 9

Rough set theory for feature selection

9.1 Introduction

This Appendix focuses on introducing the rudiments of rough set theory. It will present the basic concepts of the theory which will be illustrated by a simple tutorial example.

This Appendix is structured as follows: In Section 9.2, the decision and information systems will be described. In Section 9.3, the indiscernibility relation will be described. In Section 9.4, the lower and upper approximations will be described. In Section 9.5, the independency of attributes will be presented, and in Section 9.6, the core and the reduct concepts will be elucidated.

9.2 Decision and information systems

Data are represented as a table where each row represents an object and where each column represents an attribute that can be measured for each object. Such table is called an *"Information System"* (IS) or an *"Information Table"* (IT). To fit this definition to our research field, an information table can be seen as a representation of antigens which are defined via a set of attributes. Formally, an information system can be defined as a pair $IS = (U, A)$ where $U = \{x_1, x_2, \ldots, x_n\}$ is a non-empty, finite set of objects called the *"universe"* and $A = \{a_1, a_2, \ldots, a_k\}$ is a non-empty, finite set of *"condition"* attributes. In supervised learning, a special case of the defined information table is considered, called a *"Decision Table"* (DT) or a *"Decision System"* (DS). A DT is an information system of the form $IS = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called *"decision"*. The value set of d, called $\theta = \{d_1, d_2, \ldots, d_s\}$.

Example 2.1 *Using the terminology of RST, the data set presented in Table 9.1 can be considered as an information system $IS = (U, A)$ consisting of 4 conditional features (a, b, c, d) and 8 objects. To illustrate an example of a decision system, Table 9.2 is used. It consists of 4 conditional features (a, b, c, d), 1 decision feature (e) and 8 objects.*

Table 9.1: Information system

$x \in U$	a	b	c	d
x_1	1	0	2	2
x_2	0	1	1	1
x_3	2	0	0	1
x_4	1	1	0	2
x_5	1	0	2	0
x_6	2	2	0	1
x_7	2	1	1	1
x_8	0	1	1	0

Table 9.2: Decision system

$x \in U$	a	b	c	d	e
x_1	1	0	2	2	0
x_2	0	1	1	1	2
x_3	2	0	0	1	1
x_4	1	1	0	2	2
x_5	1	0	2	0	1
x_6	2	2	0	1	1
x_7	2	1	1	1	2
x_8	0	1	1	0	1

9.3 Indiscernibility relation

The indiscernibility relation is the mathematical basis of rough set theory. Every object of the universe is described by certain amount of information expressed by means of some attributes used for object description. Objects characterized by the same information are indiscernible in view of the available information about them. For every set of attributes $P \subset A$, an indiscernibility relation, denoted by $IND(P)$ or U/P, is defined in the following way: two objects, x_i and x_j, are indiscernible by the set of attributes P in A, if $p(x_i) = p(x_j)$ for every $p \subset P$. In other words, two objects are considered to be indiscernible or equivalent if and only if they have the same values for all attributes in the set.

The equivalence class of $IND(P)$ is called elementary set in P because it represents the smallest discernible groups of objects. For any element x_i of U, the equivalence class of x_i in relation $IND(P)$ is represented as $[x_i]_{IND(P)}$. For every object $x_j \in U$, we will use $a_i(x_j)$ to denote the value of a condition attribute a_i for an object x_j. Similarly, $d(x_j)$ is the value of the decision attribute for an object x_j. We further extend these notations for a set of attributes $P \subseteq A$, by defining $P(x_j)$ to be value tuple of attributes in P for an object x_j. The indiscernibility relation based on a subset of the condition attributes P, denoted by $IND(P)$, is defined as follows:

$$IND(P) = U/P = \{[x_j]_P | x_j \in U\} \tag{9.1}$$

$$where \ \ [x_j]_P = \{x_i | P(x_i) = P(x_j)\} \tag{9.2}$$

The indiscernibility relation based on the decision attribute d, denoted by $IND_{\{d\}}$, is defined

as follows:

$$IND_{\{d\}} = U/\{d\} = \{[x_j]_{\{d\}} | x_j \in U\} \tag{9.3}$$

Example 2.2 *In order to illustrate how a decision table from Table 9.2 defines an indiscernibility relation, we consider the following three non-empty subsets of the conditional attributes:* $\{a\}$, $\{b, c\}$ *and* $\{a, b, c\}$. *The relation* IND *may define three partitions of the universe.*

$$\begin{aligned} IND(a) &= \{\{x_1, x_4, x_5\}, \{x_2, x_8\}, \{x_3, x_6, x_7\}\} \\ IND(b, c) &= \{\{x_3\}, \{x_1, x_5\}, \{x_4\}, \{x_2, x_7, x_8\}, \{x_6\}\} \\ IND(a, b, c) &= \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} \end{aligned}$$

If we take into consideration the set a, the objects x_1, x_4 *and* x_5 *belong to the same equivalence class; they are indiscernible.*

9.4 Lower and upper approximations

The rough set approach to data analysis hinges on two basic concepts, namely the lower and upper approximations of a set, referring to the elements that doubtless belong to the set, and to the elements that possibly belong to the set. Let $P \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained by constructing the P-lower and P-upper approximations of X, denoted $\underline{P}(X)$ and $\overline{P}(X)$ respectively where:

$$\underline{P}(X) = \{x | [x]_P \subseteq X\} \tag{9.4}$$

$$\overline{P}(X) = \{x | [x]_P \cap X \neq \emptyset\} \tag{9.5}$$

Objects in $\underline{P}(X)$ can be with certainty classified as members of X on the basis of knowledge in P, while objects in $\overline{P}(X)$ can be only classified as possible members of X on the basis of knowledge in P.

Let P and Q be equivalence relations over U, then the positive, negative and boundary regions can be defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}(X) \tag{9.6}$$

$$NEG_P(Q) = U - \bigcup_{X \in U/Q} \overline{P}(X) \tag{9.7}$$

$$BND_P(Q) = \bigcup_{X \in U/Q} \overline{P}(X) - \bigcup_{X \in U/Q} \underline{P}(X) \tag{9.8}$$

The positive region contains all objects of U that can be classified to classes of U/Q using

the information in attributes P. The boundary region, $BND_P(Q)$, is the set of objects that can possibly, but not certainly, be classified in this way. The negative region, $NEG_P(Q)$, is the set of objects that cannot be classified to classes of U/Q.

Example 2.3 *Using the same decision table presented in Table 9.2, an illustrative example of the above mentioned calculations is given in what follows where* $P = \{b, c\}$ *and* $Q = \{e\}$*:*

$$\begin{aligned} POS_P(Q) &= \bigcup\{\emptyset, \{x_3, x_6\}, \{x_4\}\} = \{x_3, x_4, x_6\} \\ NEG_P(Q) &= U - \bigcup\{\{x_1, x_5\}, \{x_3, x_1, x_5, x_2, x_7, x_8, x_6\}, \{x_4, x_2, x_7, x_8\}\} = \emptyset \\ BND_P(Q) &= \bigcup\{\{x_1, x_5\}, \{x_3, x_1, x_5, x_2, x_7, x_8, x_6\}, \{x_4, x_2, x_7, x_8\}\} - \{x_3, x_4, x_6\} \\ &= \{x_1, x_2, x_5, x_7, x_8\} \end{aligned}$$

This means that objects x_3*,* x_4 *and* x_6 *can certainly be classified as belonging to a class in attribute* e*, when considering attributes* b *and* c*. The rest of the objects cannot be classified as the information that would make them discernible is absent.*

9.5 Independency of attributes

An important issue in data analysis is discovering dependencies between attributes. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted $P \Rightarrow Q$, if all attribute values from Q are uniquely determined by values of attributes from P. If there exists a functional dependency between values of Q and P, then Q depends totally on P. In rough set theory, dependency is defined in the following way: For P, $Q \subset A$, it is said that Q depends on P in a degree $k (0 \leq k \leq 1)$, denoted $P \Rightarrow kQ$, if

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \tag{9.9}$$

If $k = 1$, Q depends totally on P, if $0 < k < 1$, Q depends partially (in a degree k) on P, and if $k = 0$ then Q does not depend on P.

By calculating the change in dependency when an attribute is removed from the set of considered conditional attributes, a measure of the significance of the attribute can be obtained. The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable. More formally, given P, Q and an attribute $a \in P$:

$$\sigma_P(Q, a) = \gamma_P(Q) - \gamma_{P-\{a\}}(Q) \tag{9.10}$$

Example 2.4 *From Table 9.2, the dependency of attribute* $\{e\}$ *from the attributes* $\{b, c\}$ *is:*

$$\gamma_{\{b,c\}}(\{e\}) = \frac{|POS_{\{b,c\}}(\{e\})|}{|U|} = \frac{|\{x_3,x_4,x_6\}|}{|\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}|} = \frac{3}{8}$$

if $P = \{a, b, c\}$ *and* $Q = e$ *then*

$$\gamma_{\{a,b,c\}}(\{e\}) = \frac{|\{x_3,x_4,x_6,x_7\}}{8} = \frac{4}{8}\ ; \qquad \gamma_{\{a,b\}}(\{e\}) = \frac{|\{x_3,x_4,x_6,x_7\}}{8} = \frac{4}{8}$$
$$\gamma_{\{b,c\}}(\{e\}) = \frac{|\{x_3,x_4,x_6\}}{8} = \frac{3}{8}\ ; \qquad \gamma_{\{a,c\}}(\{e\}) = \frac{|\{x_3,x_4,x_6,x_7\}}{8} = \frac{4}{8}$$

And calculating the significance of the three attributes gives:

$$\sigma_P(Q,a) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{b,c\}}(\{e\}) = \frac{1}{8}$$
$$\sigma_P(Q,b) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,c\}}(\{e\}) = 0$$
$$\sigma_P(Q,c) = \gamma_{\{a,b,c\}}(\{e\}) - \gamma_{\{a,b\}}(\{e\}) = 0$$

From this, it follows that attribute a is indispensable, but attributes b and c can be dispensed with when considering the dependency between the decision attribute and the given individual conditional attributes.

9.6 Core and reduct of attributes

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same predictive capability of the decision feature, D, as the original. A subset $R \subseteq A$ is a reduct of A with respect to D, if R is independent and

$$\gamma_R(D) = \gamma_A(D) \tag{9.11}$$

Hence, a reduct is a set of attributes from A that preserves dependency and, consequently, set approximation. It means that a reduct is the minimal subset of attributes that enables the same classification of elements of the universe as the whole set of attributes.

The core is the intersection of all reducts. Hence, it is included in every reduct. In a sense, the core is the most important subset of attributes, for none of its elements can be removed without affecting the classification power of attributes.

Example 2.5 *Using the same decision table of Table 9.2, the dependencies for all possible subsets of A can be calculated:*

$\gamma_{\{a,b,c,d\}}(\{e\}) = \frac{8}{8}$; $\gamma_{\{a,b,c\}}(\{e\}) = \frac{4}{8}$; $\gamma_{\{a,b,d\}}(\{e\}) = \frac{8}{8}$; $\gamma_{\{a,c,d\}}(\{e\}) = \frac{8}{8}$;

$\gamma_{\{b,c,d\}}(\{e\}) = \frac{8}{8}$; $\gamma_{\{a,b\}}(\{e\}) = \frac{4}{8}$; $\gamma_{\{a,c\}}(\{e\}) = \frac{4}{8}$; $\gamma_{\{a,d\}}(\{e\}) = \frac{3}{8}$;

$\gamma_{\{b,c\}}(\{e\}) = \frac{3}{8}$; $\gamma_{\{b,d\}}(\{e\}) = \frac{8}{8}$; $\gamma_{\{c,d\}}(\{e\}) = \frac{8}{8}$; $\gamma_{\{a\}}(\{e\}) = \frac{0}{8}$;

$\gamma_{\{b\}}(\{e\}) = \frac{1}{8}$; $\gamma_{\{c\}}(\{e\}) = \frac{0}{8}$; $\gamma_{\{d\}}(\{e\}) = \frac{2}{8}$

Note that the given data set is consistent since $\gamma_{\{a,b,c,d\}}(\{e\}) = 1$. *The minimal reduct set for this example is: Reduct* $= \{\{b,d\},\{c,d\}\}$

Table 9.3: First reduct

$x \in U$	b	d	e
x_1	0	2	0
x_2	1	1	2
x_3	0	1	1
x_4	1	2	2
x_5	0	0	1
x_6	2	1	1
x_7	1	1	2
x_8	1	0	1

Table 9.4: Second reduct

$x \in U$	c	d	e
x_1	2	2	0
x_2	1	1	2
x_3	0	1	1
x_4	0	2	2
x_5	2	0	1
x_6	0	1	1
x_7	1	1	2
x_8	1	0	1

In Table 9.2, there are two possible reducts with respect to the decision attribute $\{e\}$*;* $\{b,d\}$ *and* $\{c,d\}$. *These reducts are independent with respect to the decision attribute* $\{e\}$ *and have the same dependency as the whole subset of condition attributes* A. *That means that either the attribute* b *or* c *can be eliminated from the table and consequently instead of Table 9.2, we can use either Table 9.3 or Table 9.4. The core is the attribute* d. *It is the intersection of the two possible reducts.*

9.7 Conclusion

In this Appendix, the fundamentals of rough sets for feature selection were introduced. This covered a description of the decision and information systems, the indiscernibility relation, the lower and upper approximations, the independency of attributes, and the core and the reduct concepts.

References

[AAY18] Nouman Azam, Mohammad Khan Afridi, and JingTao Yao. A game-theoretic rough set approach for handling missing data in clustering. In *Recent Trends and Future Technology in Applied Intelligence: 31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2018, Montreal, QC, Canada, June 25-28, 2018, Proceedings 31*, pages 635–647. Springer, 2018.

[AAYA18] Mohammad Khan Afridi, Nouman Azam, JingTao Yao, and Eisa Alanazi. A three-way clustering approach for handling missing data using gtrs. *International Journal of Approximate Reasoning*, 98:11–24, 2018.

[ABAB+19] David B Antcliffe, Katie L Burnham, Farah Al-Beidh, Shalini Santhakumaran, Stephen J Brett, Charles J Hinds, Deborah Ashby, Julian C Knight, and Anthony C Gordon. Transcriptomic signatures in sepsis and a differential response to steroids. from the vanish randomized trial. *American journal of respiratory and critical care medicine*, 199(8):980–986, 2019.

[ABC+16] Djillali Annane, Christian Brun Buisson, Alain Cariou, Claude Martin, Benoit Misset, Alain Renault, Blandine Lehmann, Valérie Millul, Virginie Maxime, and Eric Bellissant. Design and conduct of the activated protein c and corticosteroids for human septic shock (aprocchss) trial. *Annals of intensive care*, 6(1):1–13, 2016.

[ACE15] Kaouther Ben Ali, Zeineb Chelly, and Zied Elouedi. A new version of the dendritic cell immune algorithm based on the k-nearest neighbors. In *International Conference on Neural Information Processing*, pages 688–695. Springer, 2015.

[Alb05] Enrique Alba. *Parallel metaheuristics: a new class of algorithms*, volume 47. John Wiley & Sons, 2005.

[Alb06] Enrique Alba. *Parallel evolutionary computations*, volume 22. springer, 2006.

[ALL+18] Man Ho Au, Kaitai Liang, Joseph K. Liu, Rongxing Lu, and Jianting Ning. Privacy-preserving personal data operation on mobile cloud - chances and challenges over advanced persistent threat. *Future Generation Comp. Syst.*, 79:337–349, 2018.

[AMB21] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021.

[ARBB+18] Djillali Annane, Alain Renault, Christian Brun-Buisson, Bruno Megarbane, Jean-Pierre Quenot, Shidasp Siami, Alain Cariou, Xavier Forceville, Carole Schwebel, Claude Martin, et al. Hydrocortisone plus fludrocortisone for adults with septic shock. *New England Journal of Medicine*, 378(9):809–818, 2018.

[AT02] Enrique Alba and Marco Tomassini. Parallelism and evolutionary algorithms. *IEEE transactions on evolutionary computation*, 6(5):443–462, 2002.

[BACDBBS22] Omar Ben Amor, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. Many-objective optimization of wireless sensor network deployment. *Evolutionary Intelligence*, pages 1–17, 2022.

[BB08] Zoran Bojkovic and Bojan Bakmaz. A survey on wireless sensor networks deployment. *WSEAS Transactions on Communications*, 7(12):1172–1181, 2008.

[BBC+92] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest*, 101(6):1644–1655, 1992.

[BCP+10] Matt Bishop, Justin Cummins, Sean Peisert, Anhad Singh, Bhume Bhumiratana, Deborah Agarwal, Deborah Frincke, and Michael Hogarth. Relationships and data sanitization: A study in scarlet. In *Proceedings of the 2010 New Security Paradigms Workshop*, pages 151–164, 2010.

[BCRFPB+18] Verónica Bolón-Canedo, D Rego-Fernández, Diego Peteiro-Barral, Amparo Alonso-Betanzos, Bertha Guijarro-Berdiñas, and Noelia Sánchez-Maroño. On the scalability of feature selection methods on high-dimensional data. *Knowledge and Information Systems*, pages 1–48, 2018.

[BKRA+16] Farideh Bagherzadeh-Khiabani, Azra Ramezankhani, Fereidoun Azizi, Farzad Hadaegh, Ewout W Steyerberg, and Davood Khalili. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of clinical epidemiology*, 71:76–85, 2016.

[BL16] Ezio Bartocci and Pietro Lió. Computational modeling, formal analysis, and tools for systems biology. *PLoS computational biology*, 12(1):e1004591, 2016.

[BRdA+14] Vania Bogorny, Chiara Renso, Artur Ribeiro de Aquino, Fernando de Lucca Siqueira, and Luis Otavio Alvares. Constant–a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1):66–88, 2014.

[BS10] Chunguang Bai and Joseph Sarkis. Integrating sustainability into supplier selection with grey system and rough set methodologies. *International Journal of Production Economics*, 124(1):252–264, 2010.

[CDAB21] Zaineb Chelly Dagdia, Pavel Avdeyev, and Md Shamsuzzoha Bayzid. Biological computation and computational biology: survey, challenges, and discussion. *Artificial Intelligence Review*, 54:4169–4235, 2021.

[CDRZA20] Zaineb Chelly Dagdia, Chiara Renso, Karine Zeitouni, and Nazim Agoulmine. Towards a federated learning approach for privacy-aware analysis of semantically enriched mobility data. In *Proceedings of the 1st Workshop on Flexible Resource and Application Management on the Edge*, pages 17–20, 2020.

[CDZ21] Zaineb Chelly Dagdia and Christine Zarges. A detailed study of the distributed rough set based locality sensitive hashing feature selection technique. *Fundamenta Informaticae*, 182(2):111–179, 2021.

[CE10] Z. Chelly and Z. Elouedi. Fdcm: A fuzzy dendritic cell method. *Proceedings of the 9th International Conference on Artificial Immune Systems, Springer, ICARIS'2010*, Lecture Notes in Computer Science, 6209:102–115, 2010.

[CE11] Z. Chelly and Z. Elouedi. Further exploration of the fuzzy dendritic cell method. *Proceedings of the 10th International Conference of Artificial Immune Systems, Springer, ICARIS'2011*, Lecture Notes in Computer Science, 6825:419–432, 2011.

[CE12a] Z. Chelly and Z. Elouedi. Rc-dca:a new feature selection and signal categorization technique for the dendritic cell algorithm based on rough set theory. *Proceedings of the 11th International Conference of Artificial Immune Systems, Springer, ICARIS'2012*, Lecture Notes in Computer Science, 7597:152–165, 2012.

[CE12b] Z. Chelly and Z. Elouedi. Rst-dca: A dendritic cell algorithm based on rough set theory. *Proceedings of the 19th International Conference on Neural Information Processing, Springer, ICONIP'2012*, Lecture Notes in Computer Science, 7665:480–487, 2012.

[CE13a] Z. Chelly and Z. Elouedi. Further exploration of the hybrid fuzzy-rough dendritic cell immune classifier. *Proceedings of the 4th International Conference on E-Health and Bioengineering, IEEE, EHB'2013*, pages 1–4, 2013.

[CE13b] Z. Chelly and Z. Elouedi. A fuzzy-rough data pre-processing approach for the dendritic cell classifier. *Proceedings of the 12th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, ECSQARU'2013*, Lecture Notes in Computer Science, 7958:109–120, 2013.

[CE13c] Z. Chelly and Z. Elouedi. A new hybrid fuzzy-rough dendritic cell immune classifier. *Proceedings of the 4th International Conference on Swarm Intelligence, Springer, IC-SI'2013*, Lecture Notes in Computer Science, 7928:514–521, 2013.

[CE13d] Z. Chelly and Z. Elouedi. Qr-dca : A new rough data pre-processing approach for the dendritic cell algorithm. *Proceedings of the 11th International Conference on Adaptive and Natural Computing Algorithms, Springer, ICANNGA'2013*, Lecture Notes in Computer Science, 7824:140–150, 2013.

[CE13e] Zeineb Chelly and Zied Elouedi. Supporting fuzzy-rough sets in the dendritic cell algorithm data pre-processing phase. In *International Conference on Neural Information Processing*, pages 164–171. Springer, 2013.

[CE14a] Z. Chelly and Z. Elouedi. Improving the dendritic cell algorithm performance using fuzzy-rough set theory as a pattern discovery technique. *Proceedings of the 5th International Conference on Innovations in Bio-Inspired Computing and Applications, Springer, IBICA'2014*, Advances in Intelligent Systems and Computing, 303:23–32, 2014.

[CE14b] Z. Chelly and Z. Elouedi. A two-leveled hybrid dendritic cell algorithm under imprecise reasoning. *Proceedings of the Genetic and Evolutionary Computation Conference, ACM, GECCO'2014*, pages 97–104, 2014.

[CE14c] Zeineb Chelly and Zied Elouedi. A rough information extraction technique for the dendritic cell algorithm within imprecise circumstances. In *Hellenic Conference on Artificial Intelligence*, pages 43–56. Springer, 2014.

[CE14d] Zeineb Chelly and Zied Elouedi. A study of the data pre-processing module of the dendritic cell evolutionary algorithm. In *2014 International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 634–639. IEEE, 2014.

[CE15] Zeineb Chelly and Zied Elouedi. Hybridization schemes of the fuzzy dendritic cell immune binary classifier based on different fuzzy clustering techniques. *New Generation Computing*, 33(1):1–31, 2015.

[CE16a] Zeineb Chelly and Zied Elouedi. From the general to the specific: Inducing a novel dendritic cell algorithm from a detailed state-of-the-art review. *International Journal of Pattern Recognition and Artificial Intelligence*, 30(03):1659009, 2016.

[CE16b] Zeineb Chelly and Zied Elouedi. A survey of the dendritic cell algorithm. *Knowledge and Information Systems*, 48(3):505–535, 2016.

[CMMA+18] Francisco Javier Cabrerizo, Juan Antonio Morente-Molinera, Sergio Alonso, Witold Pedrycz, and Enrique Herrera-Viedma. Improving consensus in group decision making with intuitionistic reciprocal preference relations: A granular computing approach. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1471–1476. IEEE, 2018.

[CSE12] Z. Chelly, A. Smiti, and Z. Elouedi. Coid-fdcm: The fuzzy maintained dendritic cell classification method. *Proceedings of the 11th International Conference on Artificial Intelligence and Soft Computing, Springer, ICAISC'2012*, Lecture Notes in Computer Science, 7268:233–241, 2012.

[Dag18a] Zaineb Chelly Dagdia. A distributed dendritic cell algorithm for big data. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 103–104. ACM, 2018.

[Dag18b] Zaineb Chelly Dagdia. A scalable and distributed dendritic cell algorithm for big data classification. *Swarm and Evolutionary Computation, DOI: https://doi.org/10.1016/j.swevo.2018.08.009*, 2018.

[Dam14] Maria Luisa Damiani. Location privacy models in mobile applications: conceptual view and research directions. *GeoInformatica*, 18:819–842, 2014.

[DB22] Zaineb Chelly Dagdia and Vania Bogorny. Towards a granular computing framework for multiple aspect trajectory representation and privacy preservation: Research challenges and opportunities. In *17th CONFERENCE ON COMPUTER SCIENCE AND INTELLIGENCE SYSTEMS*, 2022.

[DD19] Melina Seedoyal Doargajudhur and Peter Dell. Impact of byod on organizational commitment: an empirical investigation. *Information Technology & People*, 32(2):246–268, 2019.

[DE18] Zaineb Chelly Dagdia and Zied Elouedi. A hybrid fuzzy maintained classification method based on dendritic cells. *Journal of Classification*, pages 1–24, 2018.

[DG00] Ivo Düntsch and Günther Gediga. Rough set data analysis. *Encyclopedia of Computer Science and Technology*, 43(28):281–301, 2000.

[DG10] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.

[dlCGSD+20] Paloma de las Cuevas, Pablo García-Sánchez, Zaineb Chelly Dagdia, María-Isabel García-Arenas, and Juan Julián Merelo Guervós. Automatic rule extraction from access rules using genetic programming. In *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pages 54–69. Springer, 2020.

[DM20] Zaineb Chelly Dagdia and Miroslav Mirchev. When evolutionary computing meets astro-and geoinformatics. In *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, pages 283–306. Elsevier, 2020.

[DMHVB13] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5, 2013.

[DZB+18] Zaineb Chelly Dagdia, Christine Zarges, Gaël Beck, Hanene Azzag, and Mustapha Lebbah. A distributed rough set theory algorithm based on locality sensitive hashing for an efficient big data pre-processing. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2597–2606. IEEE, 2018.

[DZBL17] Zaineb Chelly Dagdia, Christine Zarges, Gaël Beck, and Mustapha Lebbah. A distributed rough set theory based algorithm for an efficient big data pre-processing under the spark framework. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 911–916. IEEE, 2017.

[DZBL18a] Zaineb Chelly Dagdia, Christine Zarges, Gaël Beck, and Mustapha Lebbah. Modèle de sélection de caractéristiques pour les données massives. In *Atelier sur la Fouille de Données Complexes (FDC) — Extraction et Gestion des Connaissances, EGC 2018, Paris, France, January 23-26, 2018*, pages 1–12, 2018.

[DZBL18b] Zaineb Chelly Dagdia, Christine Zarges, Gael Beck, and Mustapha Lebbah. Nouveau modèle de sélection de caractéristiques basé sur la théorie des ensembles approximatifs pour les données massives: Méthode de sélection de caractéristiques pour les données massives. In *Conférence Internationalle sur l'Extraction et la Gestion des Connaissances*, 2018.

[DZBL20] Zaineb Chelly Dagdia, Christine Zarges, Gael Beck, and Mustapha Lebbah. A scalable and effective rough set theory based approach for big data pre-processing. *Knowledge and Information Systems*, 2020, 2020.

[DZS+18] Zaineb Chelly Dagdia, Christine Zarges, Benjamin Schannes, Martin Micalef, Lino Galiana, Benoît Rolland, Olivier de Fresnoye, and Mehdi Benchoufi. Rough set theory as a data mining technique: A case study in epidemiology and cancer incidence prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 440–455. Springer, 2018.

[FAB16] Carlos Andres Ferrero, Luis Otávio Alvares, and Vania Bogorny. Multiple aspect trajectory data analysis: research challenges and opportunities. In *GeoInfo*, pages 56–67, 2016.

[FH07] Jasmin Fisher and Thomas A Henzinger. Executable cell biology. *Nature biotechnology*, 25(11):1239, 2007.

[FLL+20] Wei Fan, Kunpeng Liu, Hao Liu, Pengyang Wang, Yong Ge, and Yanjie Fu. Autofs: Automated feature selection via diversity-aware interactive reinforcement learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1008–1013. IEEE, 2020.

[FLY+16] Zesong Fei, Bin Li, Shaoshi Yang, Chengwen Xing, Hongbin Chen, and Lajos Hanzo. A survey of multi-objective optimization in wireless sensor networks: Metrics, algorithms, and open problems. *IEEE Communications Surveys & Tutorials*, 19(1):550–586, 2016.

[Fog97] David B Fogel. The advantages of evolutionary computation. In *BCEC*, pages 1–11, 1997.

[Fre02] Alex A Freitas. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media, 2002.

[GA05] J. Greensmith and U. Aickelin. Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. *Proceedings of the 4th International Conference on Artificial Immune Systems, Springer, ICARIS'2005*, Lecture Notes in Computer Science, 3627:153–167, 2005.

[GAT06] Julie Greensmith, Uwe Aickelin, and Jamie Twycross. Articulation and clarification of the dendritic cell algorithm. In *International Conference on Artificial Immune Systems*, pages 404–417. Springer, 2006.

[GB08] Jerzy W Grzymala-Busse. Three approaches to missing attribute values: A rough set perspective. In *Data Mining: Foundations and Practice*, pages 139–152. Springer, 2008.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[GIM99] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *VLDB*, pages 518–529, 1999.

[GOGA09] F. GU, R. Oates, J. Greensmith, and U. Aickelin. Pca 4 dca: The application of principal component analysis to the dendritic cell algorithm. *Proceedings of the 9th Annual Workshop on Computational Intelligence, IEEE, UKCI'2009*, pages 1–6, 2009.

[Gre07] J. Greensmith. *The Dendritic Cell Algorithm*. Doctoral Dissertation, University of Nottingham, 2007.

[HDSZ18] Azam Hamidinekoo, Zaineb Chelly Dagdia, Zobia Suhail, and Reyer Zwiggelaar. Distributed rough set based feature selection approach to analyse deep and hand-crafted features for mammography mass classification. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2423–2432. IEEE, 2018.

[INA+16] Muhammad Iqbal, Muhammad Naeem, Alagan Anpalagan, Nadia N Qadri, and M Imran. Multi-objective optimization in sensor networks: Optimization classification, applications and solution approaches. *Computer Networks*, 99:134–161, 2016.

[JCDBBS21] Manel Jerbi, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. Malware detection using rough set based evolutionary optimization. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*, pages 634–641. Springer, 2021.

[JDBS20] Manel Jerbi, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. On the use of artificial malicious patterns for android malware detection. *Computers & Security*, 92:101743, 2020.

[JDBS22] Manel Jerbi, Zaineb Chelly Dagdia, Slim Bechikh, and Lamjed Ben Said. Android malware detection as a bi-level problem. *Computers & Security*, 121:102825, 2022.

[Kae03] Merike Kaeo. *Designing Network Security, Second Edition*. Cisco Press, 2003.

[LCA18] Xi Liu, Hanzhou Chen, and Clio Andris. trajgans: Using generative adversarial networks for geo-privacy protection of trajectory data (vision paper). In *Location privacy and security workshop*, pages 1–7, 2018.

[Len12] RG Lennon. Changing user attitudes to security in bring your own device (BYOD) & the cloud. In *Tier 2 Federation Grid, Cloud & High Performance Computing Science (RO-LCG), 2012 5th Romania*, pages 49–52. IEEE, 2012.

[LFW+19] Kunpeng Liu, Yanjie Fu, Pengfei Wang, Le Wu, Rui Bo, and Xiaolin Li. Automating feature subspace exploration via multi-agent reinforcement learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 207–215, 2019.

[LHB+13] Alba Luz León, Natalia Andrea Hoyos, Lena Isabel Barrera, Gisela De La Rosa, Rodolfo Dennis, Carmelo Dueñas, Marcela Granados, Dario Londoño, Ferney Alexander Rodríguez, Francisco José Molina, et al. Clinical course of sepsis, severe sepsis, and septic shock in a cohort of infected patients from ten colombian hospitals. *BMC infectious diseases*, 13(1):1–9, 2013.

[Lin01] Pawan Lingras. Unsupervised rough set classification using GAs. *Journal of Intelligent Information Systems*, 16(3):215–228, 2001.

[Lin02] Pawan Lingras. Rough set clustering for web mining. In *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*, volume 2, pages 1039–1044. IEEE, 2002.

[LS02] M. Lutz and G. Schuler. Immature, semi-mature and fully mature dendritic cells: which signals induce tolerance or immunity? *TRENDS in Immunology*, 23:445–449, 2002.

[Mar10] Michał Marks. A survey of multi-objective deployment in wireless sensor networks. *Journal of Telecommunications and Information technology*, pages 36–41, 2010.

[Mat01] Polly Matzinger. Essay 1: the danger model in its historical context. *Scandinavian journal of immunology*, 54(1-2):4–9, 2001.

[MBA+19] Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz Henrique Zambom Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, and Chiara Renso. Master: A multiple aspect view on trajectories. *Transactions in GIS*, 23(4):805–822, 2019.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[MO04] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

[MVH12] Keith W Miller, Jeffrey M Voas, and George F Hurlburt. Byod: Security and privacy considerations. *It Professional*, 14(5):53–55, 2012.

[MWES15] Stephen J Mooney, Daniel J Westreich, and Abdulrahman M El-Sayed. Epidemiology in the era of big data. *Epidemiology (Cambridge, Mass.)*, 26(3):390, 2015.

[NBJ11] Saket Navlakha and Ziv Bar-Joseph. Algorithms in nature: the convergence of systems biology and computational thinking. *Molecular systems biology*, 7(1):546, 2011.

[NHR+18] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, 46(4):547, 2018.

[Nur08] Paul Nurse. Life, logic and information. *Nature*, 454(7203):424, 2008.

[OSRY18] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. A non-parametric generative model for human trajectories. In *IJCAI*, volume 18, pages 3812–3817, 2018.

[PdRRG+15] Daniel Peralta, Sara del Río, Sergio Ramírez-Gallego, Isaac Triguero, Jose M Benitez, and Francisco Herrera. Evolutionary feature selection for big data classification: A mapreduce approach. *Mathematical Problems in Engineering*, 2015, 2015.

[Pri09] Corrado Priami. Algorithmic systems biology. *Communications of the ACM*, 52(5):80–88, 2009.

[RBT+21] Chiara Renso, Vania Bogorny, Konstantinos Tserpes, Stan Matwin, and Jose Antonio Fernandes de Macedo. Multiple-aspect analysis of semantic trajectories (master), 2021.

[RC14] Amir Rubinstein and Benny Chor. Computational thinking in life science education. *PLoS computational biology*, 10(11):e1003897, 2014.

[RGKH20] Jinmeng Rao, Song Gao, Yuhao Kang, and Qunying Huang. Lstm-trajgan: A deep learning approach to trajectory privacy protection. *arXiv preprint arXiv:2006.10521*, 2020.

[RJA+20] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet*, 395(10219):200–211, 2020.

[SD15] James G Shanahan and Laing Dai. Large scale distributed data science using apache spark. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2323–2324. ACM, 2015.

[SDS+16] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

[Sin12] Niharika Singh. BYOD genie is out of the bottle: "Devil or angel". *Journal of Business Management & Social Sciences Research*, 1(3):1–12, 2012.

[Tho12] Gordon Thomson. BYOD: enabling the chaos. *Network Security*, 2012(2):5–8, 2012.

[TP09] K Thangavel and A Pethalakshmi. Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing*, 9(1):1–12, 2009.

[VCR+16] Nguyen Xuan Vinh, Jeffrey Chan, Simone Romano, James Bailey, Christopher Leckie, Kotagiri Ramamohanarao, and Jian Pei. Discovering outlying aspects in large datasets. *Data Mining and Knowledge Discovery*, 30(6):1520–1555, 2016.

[WLLY21] Xingrui Wang, Xinyu Liu, Ziteng Lu, and Hanfang Yang. Large scale gps trajectory generation using map based on two stage gan. *Journal of Data Science*, 19(1):126–141, 2021.

[Woo13] Mark Woodward. *Epidemiology: study design and data analysis*. CRC press, 2013.

[WSNKK20] C Wafo Soh, LL Njilla, KK Kwiat, and CA Kamhoua. Learning quasi-identifiers for privacy-preserving exchanges: A rough set theory approach. *Granular Computing*, 5(1):71–84, 2020.

[WZWD14] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

[YJVdS19] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

[YRSF20] Chi Yunxian, Li Renjie, Zhao Shuliang, and Guo Fenghua. Correction: Measuring multi-spatiotemporal scale tourist destination popularity based on text granular computing. *PloS one*, 15(5):e0233068, 2020.

[ZGWC17] Tingting Zhai, Yang Gao, Hao Wang, and Longbing Cao. Classification of high-dimensional evolving data streams via a resource-efficient online ensemble. *Data Mining and Knowledge Discovery*, pages 1–24, 2017.

[ZTY19] Kaichun Zhou, Zongshun Tian, and Yuanwei Yang. Periodic pattern detection algorithms for personal trajectory data based on spatiotemporal multi-granularity. *IEEE Access*, 7:99683–99693, 2019.

[ZWCG17] Jianfei Zhang, Shengrui Wang, Lifei Chen, and Patrick Gallinari. Multiple bayesian discriminant functions for high-dimensional massive data classification. *Data Mining and Knowledge Discovery*, 31(2):465–501, 2017.