# UNIVERSITY OF TORINO

Research Doctorate in Modeling and Data Science

Ph.D. Thesis


# Machine Learning for Modeling the Risk of Frailty Syndrome in the Elderly Population

Supervisor: Prof. Giuseppe Costa
Co-supervisor: Prof. Mario Giacobini
Co-supervisor: Dr. Fulvio Ricceri

Candidate: Adane Nega Tarekegn

CYCLE XXXIII

October 2020

# Abstract

Frailty is a common clinical condition used to describe older people who are more vulnerable to stressors and therefore have a higher risk of negative health outcomes, such as disability, fracture, and premature death. Several definitions have been proposed in the literature to conceptualize and operationalize frailty. However, a universally accepted definition of frailty is still lacking, making it difficult to effectively target community services to older adults. Despite its challenges, frailty is not an irreversible process and can be reversed or delayed from its progression. Therefore, it is argued that it should be detected early.

This thesis focusses on early detection of frailty conditions among older adults in order to provide proactive interventions and, consequently, to maintain wellbeing and quality of life. Most early studies have focused on frailty detection by analyzing the physical performance of individuals, with a relative paucity of an administrative database. However, with the increasing number of the aging population and the growing number of frail elderly, methods to identify frailty within an administrative database are current surveillance priorities. Frailty detection using large administrative databases could capture a complex interplay of a wide variety of heterogeneous factors associated with frailty. Thus, discovering interesting patterns from such large administrative healthcare data is an important application that requires appropriate analytical tools to exploit it fully. Machine learning is a promising tool that is well suited for the analysis and capture of complex patterns within a large dataset.

This thesis presents the application of machine learning as a potential solution for modeling frailty conditions using administrative health database comprising elderly people aged 65 years and above. Both supervised and unsupervised machine learning methods have been explored to develop various models, such as for detecting and predicting adverse outcomes associated with frailty. In supervised learning, both single-label and multi-label classification methods have been examined for building frailty prediction models, while in unsupervised learning, cluster analysis is applied to identify clinically relevant clusters of complex patients. Validation of clustering results and imbalanced data classification are the most difficult problems in the machine learning paradigm. The work presented in this thesis devises new approaches for evaluating the quality of clustering results and proposes a hybrid approach for solving the imbalanced problem in multi-label learning.

Overall, the machine learning models are targeted to assist in the decision-making process aimed at achieving specific clinical health outcomes of the elderly, as well as guide the allocation of healthcare resources and reduced costs.

# Acknowledgements

I would like to express my deepest gratitude and appreciation to my supervisors, Prof. Giuseppe Costa, Prof. Mario Giacobini, Dr. Fulvio Ricceri, for their continuous support, encouragement, knowledge, and patience over the past three years. They provided me with invaluable help, suggestions, and scientific advices at every stage of my PhD study, and without their consistent inspiration, guidance, and insights, this PhD thesis would not have been possible. I also thank my supervisors for their greater support with regard to my international and research experience abroad, especially Prof. Mario Giacobni, who created the link for collaburation and helped me to organize all issues related to a short stay abroad as part of my thesis, which has also made my research experience more enjoyable. Special thank goes to Dr. Krzysztof Michalak for his support and collaborations at his institution in Poland. I studied some part of the work under his supervision, and I gratefully acknowledge the useful discussions and suggestions imparted by him.

I would also like to thank all colleagues and staff of the Unit of Epidemiology Regional Health Service (SEPI) research group in Grugliasco, Torino, for their great support and warm friendship. They were always open and supportive whenever I have inquires, especially Dr. Fulvio Ricceri, who made a tremendous contribution to the office and data resources at SEPI and guiding me through this hard but exciting journey. I would also thank Ms. Elisa Ferracin for her assistance in understanding the data of this project well.

I would like to express my special thanks to the University of Torino for awarding me a Ph.D. scholarship, which provided me the funding sources that sustained me during my study. In connection with this, I would thank the members of the Ph.D. committee, especially to the Ph.D. coordinator, Prof. Laura Sacerdote, for her valuable support and guidance linked to my Ph.D. experience. She has helped me to solve all the issues related to my

study both from the administrative and academic point of view.

Furthermore, I would like to thank my colleagues of the Ph.D. program in modeling and data science for their great support and warm friendship. I would also like to thank Irene Azzali, at the department of veterinary sciences, for her great help with the conceptual and technical skills for genetic programming.

Last but not least, a very special thanks must go to my parents and families. Your continuous support and encouragement helped me towards my whole academic path. Thank you.

# Publications

Part of the work within this thesis have been documented in the following publications:

1. Adane Nega Tarekegn , Fulvio Ricceri , Giuseppe Costa, Elisa Ferracin, Mario Giacobini. "Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches". *JMIR Medical Informatics*. Vol.8, No.6, June 2020.

2. Adane Nega Tarekegn, Fulvio Ricceri, Giuseppe Costa, Elisa Ferracin, Mario Giacobini. "Detection of Frailty Using Genetic Programming". *23rd European Conference on Genetic Programming (EuroGP): Lecture Notes in Computer Science*, Seville, Spain: Springer, p.228–243, Vol.12101, April 2020.

3. Adane Nega Tarekegn, Krzysztof Michalak, Mario Giacobini. "Cross-Validation Approach to Evaluate Clustering Algorithms: An Experimental Study using Multi-label Datasets". *SN Computer Science*.Vol.1, issue 5,No.263, September 2020.

4. Adane Nega Tarekegn, Mario Giacobini, Krzysztof Michalak. "A Review of Methods for Imbalanced Multi-Label Classification". *Pattern Recognition*. (Under review)

5. Adane Nega Tarekegn, Krzysztof Michalak, Fulvio Ricceri, Giuseppe Costa, Mario Giacobini. "Learning from Imbalanced Multi-label Data for Early Detection of Frailty Syndrome". *Artificial Intelligence in Medicine*. (Under review)

# Contents

# List of Tables

# List of Figures

:

# List of Abbreviations

| | |
|---|---|
| ANN | Artificial neural network |
| BR | Binary relevance |
| Card | Cardinality |
| CC | Classifier chains |
| CLR | Calibrated label ranking |
| CVIR | Coefficient of variation of IRLbl |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| Dens | Density |
| DT | Decision tree |
| ED | Emergency department |
| EM | Expectation-Maximization |
| GP | Genetic programming |
| HDBSCAN | Hierarchical DBSCAN |
| IRLbl | Imbalance ratio per label |
| KNN | K nearest neighbours |
| LP | Label powerset |
| LR | Logistic Regression |
| MaxIR | Maximum imbalance ratio |
| MCA | Multiple correspondence analysis |
| MeanIR | Mean imbalance ratio |
| ML | Machine learning |

:

| | |
|---|---|
| MLC | Multi-label classification |
| MLD | Multi-label dataset |
| MLKNN | Multi-Label k Nearest Neighbors |
| MLR | Multi-label ranking |
| ML-TLSMOTE | Multi-label SMOTE with Tomek links |
| MMP | multi-class multi-label perceptron |
| PCA | Principal Component Analysis |
| PPT | Pruned problem transformation |
| RAkEL | Random k-label powerset |
| RF | Random forest |
| ROC | Receiver Operating Characteristics |
| RPC | Ranking by pairwise comparison |
| SMOTE | Synthetic minority oversampling technique |
| SOMS | Self-organizing maps |
| SVM | support vector machine |
| t-SNE | t-Distributed Stochastic Neighbour Embedding |
| WHO | World Health Organization |

# Chapter 1

# General Introduction

## 1.1 Background and Motivation

Human life expectancy has increased markedly throughout the world during the past 100 years, due to improvements in survival that occur during the demographic change [1]. This demographic transformation of the population has resulted in a much higher proportion of people living in advanced old age. The trend is accelerating quickly, and projections foresee a growing number and share of older adults (aged 65 years and over), with a particularly rapid increase in the number of very old people (aged 85 years and over)[1, 2]. Between 2000 and 2050, the number of older adults aged over 80 years is expected to increase fourfold [3]. While this increased longevity may be seen as a remarkable human success story, which represents the achievement of public health, medical advancement, and economic development over disease and injury [4], it also presents tremendous challenges and consequences. One of the most problematic consequences associated with the aging of the population is the clinical condition of frailty. Frailty has become the focus of considerable scientific research over the past 18 years, aiming to evaluate the health status of older adults and a need to prevent or at least delay the onset of the late-life disability and its adverse consequences [5]. However, no formal agreement has yet been reached on the comprehensive and objective definition of frailty. That is, the concept of frailty remains controversial, despite its significant impact on individuals and society with increased risk of dependency, disability, hospitalization, and mortality [6, 7].

In spite of all these conceptual disagreements and disputes on how to assess frailty, most studies recognize it as a common clinical syndrome among

older people, which reflects the state of decreased resilience and increasing vulnerability to stressor events (such as acute illness, surgery or trauma), along with a higher risk of adverse health outcomes [8]. The level of frailty can range from none frail to advanced frail and appears to be a dynamic state in which people can be less frail or more frail over time [9]. Frail people are considered to be a group of patients that have the most complex and challenging health problems. They can also have multiple chronic conditions simultaneously, which increases the complexity of their healthcare needs [10]. Globally, one of the important challenges to healthcare is the considerable growth in the proportion of frail older people with often complex needs and increased utilization of healthcare resources [11, 12]. The complexity of frailty is also influenced by a large number of factors (biological, genetic, social, environmental, etc.) and consequently, elderly patients are a heterogeneous group in that their frailty condition may involve multidimensional functional losses (cognitive, physical, social, psychological) [13, 14]. Hence, the care of frail elderly adults is challenging due to complex comorbidities with multidimensional deterioration, compounded by the need for consistent, ongoing management in spite of fragmented healthcare service delivery [15]. It is, therefore, argued that early identification and intervention is an important public health goal in order to prevent or delay the adverse effect of frailty while maintaining life expectancy. Notwithstanding the burden of frailty, there is increasing evidence indicating that frailty can be delayed or reversed [16, 17], and the attention has turned to challenge for better identifying older people at risk of frailty with the aim of improving health outcomes and quality of life. The early identification and diagnosis of pre-frail and frail older adults through various screening methods can provide an opportunity to effectively target interventions to better manage frailty and improve health and wellbeing [18, 19].

Several landmark studies have been extensively applied for frailty that demonstrate its association with adverse health outcomes. Most of them are either questionnaire-based methods or analysis of the physical performance of the subject. Among those, the Fried's phenotype model [20] and the Rockwood's accumulation of deficits model [21] are some of the most commonly used frailty measurements. Even though they are simple, these measures are not recorded in the current administrative databases and can be impractical when considering a large fraction of the population [22]. Therefore, with the increasing number of an aging population and the growing number of frail elderly, methods to identify frailty within administrative

databases (patient and population levels) are current surveillance priorities [23]. Besides, such identification of frailty using administrative database could capture the current views on frailty through a complex interplay of a wide variety of heterogeneous factors and includes more than its physical dimensions, also integrating its cognitive, psychological, and social components, in order to reflect the multidimensional and multisystem impairments and consequences that are inherent to the frailty syndrome [24].

This thesis uses an administrative health database, which contains about one million older adults aged 65 years and above, representing the whole elderly population of the Piedmont region, Italy. Healthcare administrative database represents a powerful alternative that can offer better, more robust and flexible solutions for a large population and also uncover information that is unlikely missed in a questionnaire or an interview due to social desirability bias [25]. These days, electronic medical administrative data is often generated from the healthcare domain in large quantities, requiring appropriate analytical tools to fully exploit it and discover interesting patterns. Machine learning is an exciting and promising tool that is well suited for the analysis of such a large amount of healthcare administrative databases. Machine learning can offer a potential solution to improve efforts at identification and prediction of frailty in older adults, particularly with regard to the use of administrative datasets that are often characterized by large dimensionality (high number of features), class imbalance distribution (e.g., many more healthy patients than sick), a vast number of samples, information collected from different sources (e.g., clinical examination, demographic and socioeconomic sources), etc.

Machine learning is a branch of artificial intelligence that offers classification, prediction, and clustering capabilities focused on building automated systems that support clinical decision making and visualization of information for shared decision making [26]. Machine learning algorithms analyze the data and generate solutions (models) to address a large variety of complex problems. Once the model is inferred, if we want to understand how the algorithm performed certain choices, or we interpret the structure of the solutions, we have the possibility to learn the hidden structure of the data. Machine learning has evolved rapidly over time, resulting in a revolution in healthcare [27] that includes a wide variety of sophisticated and new computational methods for enhanced data analysis. In the aging population, it can be applied from frailty identification through remote health monitoring of elderly people. Despite the advantageous position of machine

learning methods and multi-label learning, more specifically, to aid identification, prediction, and treatment of frailty within administrative databases, there has only been limited application within the published research literature to date. One early research compared the machine learning algorithm with a self-reported Frailty Index (FI) to predict survival within a community sample of older patients [28], showing that the machine learning algorithm outperformed the FI. Other subsequent researches have focused on the use of machine learning techniques to identify frailty by extracting data from electronic clinical notes of various client populations, including patients with heart failure [29], Medicare patients [30] and health care insurance enrollees [31], with another recent study supplementing electronic health records with interview data among older community center attendees [32]. More recently, prediction of hospital admission for older adults [33] was developed using machine learning techniques based on contact assessment data, including a series of geriatric syndromes, functional assessments, and baseline care needs. To the best of the author's knowledge, however, there are no adequate studies that have applied machine learning algorithms to simultaneously predict multiple outcomes associated with frailty within the population-based administrative database, nor attempted to compare the performance of several commonly used supervised learning algorithms, particularly using multi-label learning approaches.

This thesis investigates the use of machine learning, with its rich knowledge representations, as an alternative solution to identify an individual or group of individuals with frailty (i.e. building models that are able to discriminate between frail and non-frail people) and to predict future adverse outcomes associated with frailty condition based on electronic administrative databases. Both supervised and unsupervised machine learning methods have been explored to develop various models for the assessment and management of frailty in elderly people aged 65 years and above. In supervised learning, both single-label (for one outcome prediction) and multi-label classification (for multiple outcomes prediction simultaneously) models have been developed, while in the unsupervised learning, clustering analysis and its validation has been explored for analyzing the frailty dataset aiming at estimating the number of clusters and assessing the prediction strength of models to assign cluster labels for new samples.

However, with the use of machine learning techniques, several challenges exist which prohibits the efficient development of predictive models. In

supervised learning, one of the open challenges is the presence of highly imbalanced data, particularly in a multi-label classification. This thesis first identifies the gap through conducting a systematic review on the state-of-the-art methods, and then proposes an alternative approach for addressing imbalance problem in multi-label classification, in the specific context of frailty data. In unsupervised learning, determining the quality of the results obtained by clustering techniques is a key issue. However, very little practical guidance is available to measure the quality of the clustering algorithms in assigning unknown observations to clustering groups. The thesis also investigates the development of a new clustering validity measure that helps assessing the clustering results.

In general, the machine learning models are designed to assist in the decision-making process aimed at achieving specific clinical outcomes of frailty, as well as guide the allocation of healthcare resources and reduced costs.

## 1.2 Aims and Scope

This thesis is an investigation of how machine learning techniques can be utilized for the identification and prediction of frailty conditions using an administrative medical database. More specifically, the objectives in this thesis are: (1) to develop and evaluate clinical classification models (using both single-label and multi-label classification algorithms) for predicting negative health outcomes associated with frailty; (2) to explore the potential of clustering algorithms for identifying subgroups of complex patients from a large sample of the elderly population; (3) to examine and propose a new cluster validation criteria for evaluating clustering results in a quantitative manner and a hybrid approach for handling an imbalanced problem in multi-label classification. More detailed information on each specific objectives are given below:

1. **To target interventions on those who will become frail or those who are at high risk of being frail**.
   Identifying high-risk individuals is often perceived as an important part of prevention programs [34], since the available healthcare for older adults may be insufficient, or may not be designed to meet their needs. From this perspective, the prediction of older persons at risk of admission to hospital or other negative outcomes may be one important way

for the future healthcare system to act proactively when meeting increasing needs for care. Here, single-label supervised machine learning techniques can be used, which are particularly suited for identification and predictions of frailty based on an administrative database. Various types of commonly used machine learning models have been explored and compared to detect people at an early pre-frail stage (i.e., those at high risk of becoming frail) and those who are already frail and therefore at risk of complications of frailty (such as disability, falls, hospitalizations and mortality). Therefore, we wanted to develop and evaluate clinically useful models for detecting and predicting adverse outcomes of older persons based on routine healthcare administrative databases.

2. **To identify subgroups of the target population in the elderly using cluster analysis**.
Unsupervised learning methods, such as clustering that uses large administrative data can help to group and organize the elderly population into distinct categories, each with specific needs, characteristics, or behaviors to allow care delivery and services to be tailored for these groups [35]. While it is practically impossible to develop models and intervention programs for each individual, programs can be created for groups of patients with largely similar characteristics. The exponential growth in medical datasets, together with the developments in machine learning tools, provides new opportunities to use data for cluster analysis [36]. Specifically, in this thesis, the clusters identified using frailty data can be used to prioritize interventions among the pre-frail, mild frail, and advanced frail groups, and there would also be the potential to set financial budgets for individuals per cluster.

3. **To build a multi-label predictive model for early diagnosis and prediction of simultaneous outcomes associated with frailty**.
People with frailty may be diagnosed with more than one health condition, and because of the high prevalence of frailty and its multiple adverse conditions, it is often more important to target several outcomes simultaneously than a single outcome. Simultaneous prediction of multiple outcomes of frailty from a common set of predictors can help to detect multiple complicated issues in older adults, minimizing the risk of multiple states at a time. Recently, multi-label learning is used to handle the task of simultaneous prediction of all target outcomes.

Following the objectives in steps 1-3, this thesis also proposes new solutions for two important problems in machine learning: (1) Cluster validation criteria for assessing the quality of clustering results, and (2) A hybrid resampling approach for imbalanced data classification in multi-label learning.

## 1.3 Thesis Contributions

The primary contribution of this thesis is the efficient development of predictive models, their application for managing and preventing the progression of frailty condition in older adults. Motivated by the results obtained in this part, the second contribution is the introduction of a new method for addressing the issue of cluster validation. This method uses the structure of multi-label datasets to elaborate a novel clustering validation criteria. The proposed validation index is a simple yet effective label-based approach to measuring the predictive strength of a clustering model. This validation criterion has been used for evaluating the clustering results that have been obtained on the dataset of the older population. The third contribution of the thesis is dedicated to addressing an imbalanced problem in multi-label classification, aiming to provide simultaneous predictions of multiple outcomes associated with frailty. A hybrid of SMOTE and T-link approaches is proposd to address the intrinsic problem of imbalanced data in the multi-label context. This hybrid method particularly solves the imbalanced label distributions (the joint occurrence of minority and majority labels that appear in the same instances) in a multi-label classification. Despite the advantages of such a hybrid method for imbalanced classification, it has not been well grounded in the multi-label scenario. The proposed method was evaluated considering the imbalanced multi-label administrative dataset of older adults aged 65 years and above.

## 1.4 Thesis Organization

This thesis is organized into seven chapters: the introduction, the preliminary chapter that provides the context, and the background of the material on which the thesis is built on. The overall structure of the thesis, excluding this introductory chapter, is organized as follows:

Chapter 2 – Literature Review. This chapter introduces the key concepts necessary to understand the content of the thesis fully. It presents the review of background information on the concepts of frailty, aging and frailty, impacts and healthcare interventions of frailty, machine learning methods, and application of machine learning in the healthcare domain.

Chapter 3 – Imbalanced Data classification: A systematic review. This chapter presents the first survey of handling the imbalanced problem in multi-label data classification (MLC) which includes a comprehensive survey of the state-of-the-art methods for imbalanced MLC, the characteristics of the data, problem descriptions, solutions and limitations of the approaches proposed in the existing literature for solving an imbalanced problem.

Chapter 4 – Predictive Modelling for Frailty Conditions in Older Adults. A detailed description of the dataset used (i.e., the frailty data), data preprocessing, feature selection, imbalanced dataset, and the classification model development is provided. It also presents the data analysis used in the study, the evaluation methodologies, and the discussion of results.

Chapter 5 – Multi-label Classification for Early Diagnosis of Frailty Syndrome. It introduces the use of multi-label learning for early diagnosis and prediction of adverse outcomes associated with frailty. This chapter discusses the variety of multi-label classification models, evaluation metrics, and experimental results. It also presents an algorithm proposed for solving the imbalanced problem in multi-label classification.

Chapter 6 – Cluster Analysis and Its Validation: Application for improving health Conditions of Elderly. This chapter is split into two parts. The first part focuses on proposing a new approach to cluster validation, investigating the use of a cross-validation procedure to evaluate the quality of the clustering results. The second part covers the application of cluster analysis for identifying subgroups of patients in the elderly. It describes and applies clustering techniques to the administrative dataset. It also presents the dimensionality reduction techniques used, the variety of clustering methods, and a methodology proposed for the evaluation of clustering algorithms.

Chapter 7 – Conclusions and Future Work. This chapter summarizes the results and the main research objectives achieved in each chapter of the

thesis. The chapter also provides further discussions on more general topics covered in the whole thesis and the possible areas of future work.

# Chapter 2

# Literature Review

This chapter reviews the literature related to this thesis. It discusses an overview of frailty and of machine learning techniques. The first section of this chapter (section 2.1) introduces the topic of frailty in older adults and summarizes its current conceptualizations, impacts from societal and psychological perspectives, and intervention methods. Then section 2.2 presents the main concepts behind each step of the model development process in this thesis. It provides an introduction of machine learning methods with a particular focus on supervised and unsupervised learning paradigms. Finally, section 2.3 describes the application of machine learning in the healthcare domain with some examples of applications to enhance clinical decision making.

## 2.1 Introduction to Frailty

The notion of frailty has evolved substantially over the past two decades in the context of population aging, and the number of publications on frailty syndrome has increased exponentially [37]. Yet, the concept of frailty and its objective diagnostic parameters are still evolving as a complex phenomenon. Frailty has been viewed as a cornerstone of geriatric medicine and has been described as the most problematic expression of population aging [38]. It has been shown that frailty has become a major challenge as population ages. The age composition of the world population is changing, and the relative number of older adults that grow old is increasing quickly, due to the increased life expectancy and the decreased fertility rates [39]. Because of this rapid growth of the older population, national health, and social

care budgets are under pressure. Healthcare systems have encountered significant challenges urging innovation in the management of frail elderly. In developed countries, the estimated prevalence of frailty is around 11% for people aged above 65 years, rising to 25–50% for people above 85 years old [40].

Current literature shows that detecting (recognizing) frailty among older populations at an early stage provides options to enable proactive intervention and helps to reduce the burden of adverse outcomes associated with frailty, consequently, to support independent living [41].

### 2.1.1 Definition of Frailty

Frailty is a term used by healthcare professionals, policymakers, researchers, and laypeople to describe a range of conditions in older people. However, their understanding and usages of the term are not similar, which clearly shows the difficulty of the issue at hand. Despite a large number of scales and questionnaires to detect frailty syndrome, there is no universally agreed operational definition of frailty or a generalized method to diagnose or for its screening, as groups of clinicians and researchers have demonstrated disparate views on the characteristics that make up frailty. The lack of a gold standard definition of frailty causes heterogeneity in studies. Nevertheless, most definitions of conceptual frailty contain some common factors, including increased vulnerability to adverse health outcomes, decreased capacity/reserve to adapt minor stressor events, and impairment in various physiological systems [9]. With the inclusion of these factors, WHO has defined conceptual frailty as " a clinically recognizable state in which the ability of the older people to cope with everyday or acute stressor is compromised by an increased vulnerability brought by age-associated declines in physiological reserve and function across multiple organ systems" [42].

Considering the composition and type of variables of frailty scores, four main definitions of frailty have been distinguished in the literature. The first is the **"phenotype of frailty"** approach, which considers frailty as 'a biological syndrome of decreased reserve and resistance to stressors resulting from decline across multiple physiological systems and causing vulnerability to adverse outcomes'[20]. This definition is the most widely adopted one, which is based on physical markers (such as global weakness with low muscle strength), overall slowness (such as slow gait speed), reduced balance and mobility, exhaustion, involuntary weight loss, and low physical activity. For screening the presence of frailty, at least three of these symptoms must

be observed. The presence of one or two of the symptoms indicates pre-frailty. However, this definition has been criticized as insufficient by other researchers as it doesn't consider socioeconomic domains (such as living alone, income), cognitive and mental health domains to define frailty. Second, the **"accumulation of deficit"** approach counts the number of health problems or deficits to classify an individual as frail [43]. It takes into account various baseline parameters of signs, symptoms, abnormal laboratory values, disabilities, and disease states, collectively named as deficits, to define frailty. The acquired frailty index is a computation of the presence or absence of each variable as a proportion of the total; thus, frailty is defined as the cumulative effect of individual deficits -'the more individuals have wrong with them, the more likely they are to be frail'. However, the frailty index does not distinguish frailty from disability or comorbidity. Third, the **"multidimensional"** approach defines frailty as a dynamic process of loss of function in one or more domains, including physiological, psychological, cognition, nutrition, and social domains, making the individual more vulnerable [9]. Fourth, the authors in [44] have proposed a **"disability"** approach to define frailty, thinking that primarily frailty scores are created with variables representing a degree of disability. They included this definition of frailty without a reference from literature or without any theoretical basis, even though disability is considered as an outcome of frailty by several authors [45]. Whereas frailty indicates to instability and risk of loss of function, disability refers to the loss of function and often assessed based on dependency or difficulty in performing activities necessary to live independently, such as activities of daily living (ADL), e.g., dressing, bathing, eating, toileting, continence, and transferring [46] and instrumental activities of daily living (IADL), e.g., shopping, meal preparation, telephone use, laundry, housekeeping, medication, transportation, and finances [47].

Different models of frailty give varied prevalence estimates that constitute a major challenge in comparing the results across different studies [48]. In spite of the differing frailty conceptualizations, its dynamic nature can be of considerable importance to the development of interventions for prevention and treatment.

### 2.1.2 Frailty in Older Adults

The number of older adults has increased significantly in most current societies. With an increasingly aging population, there will be a greater prevalence of frailty [13]. Its prevalence varies greatly according to the age of the

population under study and is often higher in women than men. Frailty is a progressive age-related decline in multiple physiological systems, which collectively results in a vulnerability to sudden health status changes triggered by relatively minor stressor events. Frailty is often associated with individuals who are functionally dependent on others for activities of daily life [38]. Aging is associated with a gradual decline in physical functioning, and the condition of frailty is considered as an increasingly problematic consequence of aging [13]. However, the rate of physical decline and functional status of older individuals vary significantly, depending on their biological, genetic, and environmental backgrounds as well as other physical, psychological, and social factors. As a consequence, elderly patients are heterogeneous groups in which the expression of frailty may include comorbidities and the loss of multidimensional functions (cognitive, psychological, physical, social) that may need a broad array of healthcare services. Hence, individuals with the same chronological age can have different biological ages [49], and advanced aging on its own does not necessarily couple with frailty [50]. The reasons are two: first, compared to other older adults, frail older adults show a degradation of multiple physiological systems that are responsible for healthy adaptation to the different demands of life [51]. Those older adults considered frail are particularly vulnerable to undesirable outcomes, such as hospitalization, institutionalization, injurious falls, or death [8]. Thus, frailty is a better indicator for those in need of interventions for their health and wellbeing than chronological age [52]. Second, frailty is at least partly programmed in early life and is also associated with lower socioeconomic status in adulthood [53]. Thus, frailty is related to and also distinct from the natural aging, represents those adults at the greatest risk of adverse outcomes and is heterogeneous [54].

### 2.1.3 Impacts and Interventions of Frailty

Frailty represents a huge public health issue at both the patient and the societal levels because of its multiple clinical, economic, and societal consequences [55].

The ecological perspective considers the social impact of frailty at many levels, from the individual, through expanding spheres of family, friends and caregivers, peer groups, institutions, neighborhoods, and communities, to society at large [56]. From a clinical perspective, frailty is crucial because people with frailty are at increased risk of premature death [57], and various negative health outcomes, including fracture, disability, falls, depression and

dementia, all of which result in a poor quality of life, increased cost and use of healthcare resources, such as hospitalization, emergency department and institutionalization [7]. From a recent systematic review [58], it has been shown that the healthcare costs of frail older people are sometimes much greater than those of the non-frail counterparts. It has also been observed that healthcare costs are increasing all over the world, promoting healthcare services to seek reduced hospital admission and readmission rates, shorter lengths of stay, and postponement of admission to nursing homes [59, 60]. This is due to the fact that current healthcare services are mostly designed to address disease-specific and organ-specific problems one at a time and are not well prepared to deal with the chronic and medical needs of frail older patients and to provide seamless care for them in the long term [61].

Frailty is important from a societal perspective because it identifies groups of people in need of extra hours of home care services, medical attention, and at risk of higher dependency, particularly when functional dependence increases an individual's reliance on caregivers. On the other hand, the more hours of care provided, the greater risk of health and economic consequences to the caregiver [62]. Frailty is also a concern when considering financial health care planning to better select management and prevention programs.

Reducing the prevalence or severity of frailty with appropriate interventions has large benefits for individuals, their families, and society. Several interventions for frailty have been proposed to improve health outcomes of frail individuals, including exercises interventions [63], nutritional intervention [64], multicomponent interventions [65], pharmacological agents [66] and individually tailored geriatric care models based on comprehensive geriatric assessment (CGA) [67]. Therefore, identifying frail older subjects who are frail from people who are not frail should be an essential aspect of assessment in any healthcare system. Early diagnosis or detection of frailty conditions can help to improve care for elderly people, reducing the risk of pre-frail states progressing into frail states (primary prevention). Early detection can also be important for implementing therapeutic measures, which can help to decrease or delay the underlying symptoms and conditions or enhance the impacts on independence or a healthy lifestyle (secondary prevention). In more advanced stages, frailty diagnosis provides valuable information, essential for planning and implementing intervention strategies aimed to preserve functional status or control the progression of adverse outcomes, such as institutionalization, recurrent hospitalizations,

or death (tertiary prevention)[68]. The evidence from the study of various types of frailty interventions demonstrates that frailty can be managed and reduced. Frailty screening can also provide information on older people at high risk of disability and poor prognosis, and help to identify reversible risk factors. In the presence of different clinical scenarios where the care of the elderly is a priority, such as nursing homes or primary health care, it is imperative to have a specific model for the detection of frailty according to the characteristics of the population being studied.

## 2.2 Introduction to Machine Learning

### 2.2.1 Machine Learning Paradigms

Machine learning (ML) is a broad and rapidly growing area of research in the field of artificial intelligence [69]. The definition of what exactly constitutes machine learning has been varied among the different experts in the field. In 1959, Arthur Samuel [70] defined machine learning as "a subfield of computer science that gives computers the ability to learn without being explicitly programmed." It means that ML is able to perform a specified task without being directly told how to do it. The other most widely accepted formal definition of ML given by Tom Mitchell [71] was as follows:

**Definition 2.2.1.** *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E .*

In short, Mitchell defines ML as "a set of computer programs that automatically learn from experience." According to this definition, we can formulate our frailty problem as the task of predicting older adults who are at increased risk of frailty (*Task T*) using relevant clinical, administrative data from several individuals (*experience E*) through an ML algorithm. If the algorithm has successfully learned (*measure P*), it will then be capable of using these data to predict clinical outcomes of new older individuals.

In the field of today's data science, ML aims to select, explore and extract useful knowledge from complex, often non-linear data, building a computational model capable of describing unknown patterns or correlations, and in turn, solve challenging problems. This learning process is often carried out through repeated exposure to the defined problem (dataset), allowing

the model to achieve self-optimization and continuously enhance its ability to solve new, previously unseen problems [72]. ML draws on concepts from many different fields, including computer science, statistics, and optimization. At their core, almost all ML problems can be formulated as an optimization problem with respect to a dataset. In such settings, the goal is to find a model that best explains the data. Often, ML is related to pattern recognition, artificial intelligence, data mining, and statistics.

Figure 2.1 is an indicative representation of the degree of similarities and differences among the various fields. Although there are no clear boundaries among these areas and they usually overlap, there are some basic differences between machine learning and statistics:

1. ML focuses on the task of prediction, by using general-purpose learning algorithms to find patterns usually in a massive dataset.On the other hand, statistical methods have a long-standing focus on inference, which is attained through the creation and fitting of a problem-specific probability model [73].

2. Most ML techniques are hypothesis-free (i.e., no prior assumptions are required about the underlying distribution of data), as the goal is to reconstruct associations within the dataset, while traditional statistics often depends on specific hypotheses and assumptions, which are usually originated from the model that has produced the dataset [74].

3. The metrics used to evaluate the generalization performance of an ML model (ROC curve, cross-validation, etc.) are generally distinct from the statistical approach, which primarily relies on the computation of the P values to accept or reject a null hypothesis [75, 76].

4. Statistical modeling is generally fitting a parsimonious model to produce an easy to understand and interpretable results. However, biological and clinical factors are often dependent on each other, and their relationship may be non-linear. ML approach, however, considers all possible associations between features with complex non-linear patterns, while investigating to capture as many informative and interesting features as possible, which may produce a complicated and sophisticated model that is not easy to understand or interpret.

In general, statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns. Both inference and prediction are major goals in the study of biological systems. Inference

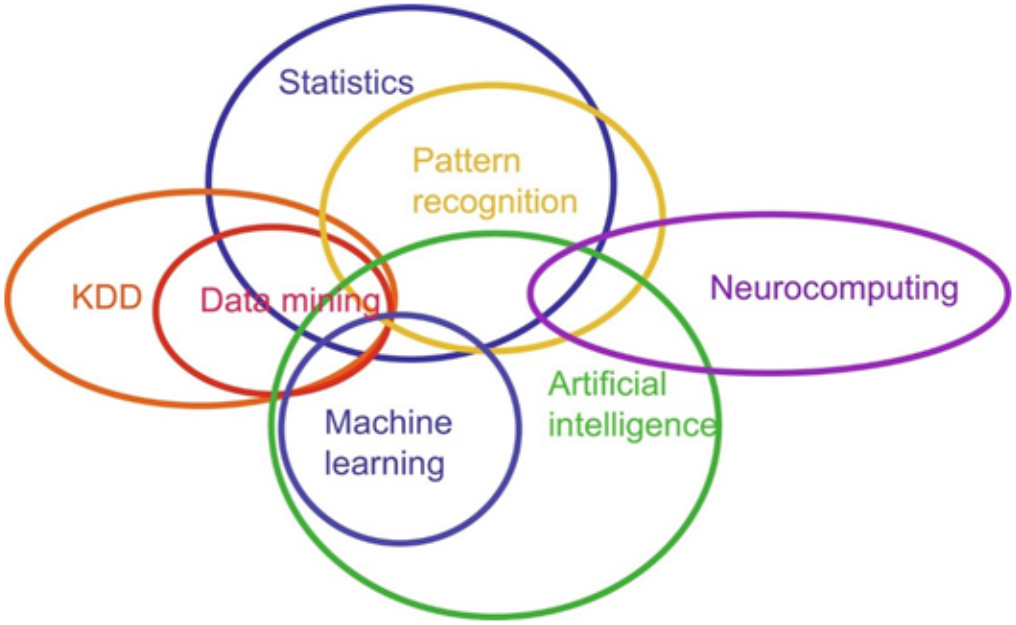


Figure 2.1: Representation of degree of similarities between various fields

creates a mathematical model of the data generation process to formalize understanding or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future behavior that makes it possible to identify best courses of action (e.g., treatment choice) without requiring understanding of the underlying mechanisms. In a typical research project, both inference and prediction can be of value—we want to know how biological processes work and what will happen next. Therefore, these two approaches can be integrated in a way that they can determine an added value in medical care.

There are many different applications of ML, which fall into one of four broad categories: supervised, unsupervised, semi-supervised and reinforcement learning (Figure 2.2).
**Supervised learning**: is used to search the relationship between input variables (a set of features) and one or more output variables (classes or labels) and then forms a function that predicts the outcome value for a set of unlabeled samples based on an acceptable degree of performance [76]. In supervised learning, the training dataset should have the correct input-output pairs and has two major tasks to be performed: classification, where the task is to predict the class or group to which a new sample should be

assigned, and regression, where the values of a continuous variable for a new sample must be estimated.

**Unsupervised learning**: attempts to find the structure in the data without the need for training data, labels, or classes. It explores the underlying structure of the data to identify useful patterns, such as clusters. The main tasks in unsupervised learning include cluster analysis [77], dimensionality reduction [78], self-organizing maps (SOM) [79], representation learning [80], and density estimation [81].

Supervised and unsupervised learning are by far the most commonly used approaches in general, and the work presented in this thesis employs these two learning approaches and, therefore, will be discussed in greater detail in the next subsections.

**Semi-supervised learning**: is a combination of the two previous paradigms where the target variables or classes are available for only a part of the data. The aim of semi-supervised learning is to classify the available unlabeled data set using its labeled information set. In such scenarios, the size of the unlabeled data should be greater than the labeled set. Otherwise, the learning problem can be addressed through supervised learning.

**Reinforcement learning**: is a paradigm where the algorithms' behavior (i.e., the learner) is shaped through a sequence of rewards and penalties, which is based on the actions it takes in the environment towards a defined goal. Unlike supervised learning, where the algorithm uses a set of examples to model behavior, the learning algorithm in reinforcement learning is allowed to behave freely, i.e., on the basis of trial and error, to discover what actions maximize reward and minimize the penalty. A computer chess-playing with a human is an example of reinforcement learning.

### 2.2.2 Supervised Learning

In this learning paradigm, the algorithm knows the output variable that it is trying to predict, i.e., the target variable; this could be, for instance, the presence or absence of disease, severity of symptoms, or future clinical outcome. The goal is to use an algorithm to learn the optimal function that best captures the relationship between the input and output variables. The algorithm is trained using several samples and is allowed to receive feedback during the learning process based on how close its prediction matches the true value. Depending on whether the target variable is a categorical or continuous variable, the supervised learning task is either a regression or classification problem, respectively.

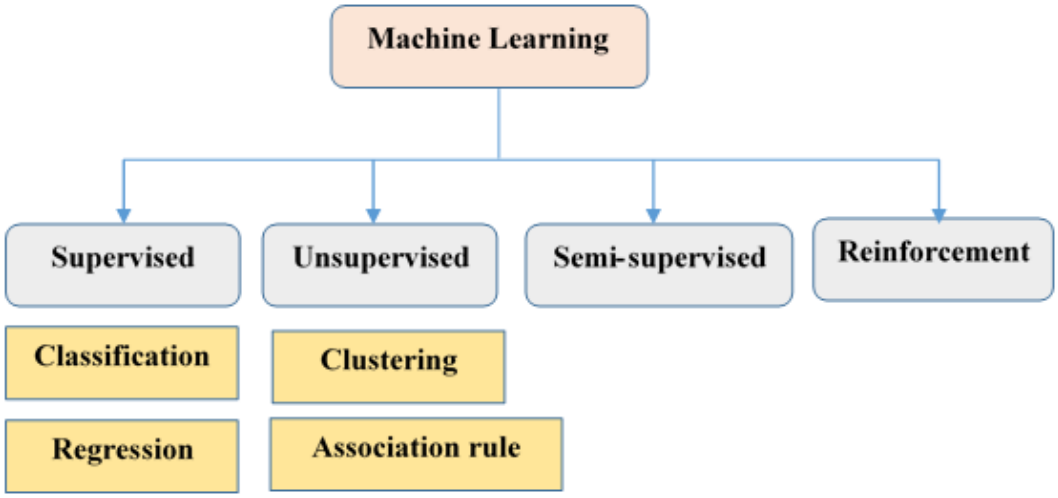


Figure 2.2: Categories of learning paradigms in machine learning

In a regression problem, the goal is to predict a numeric score on a continuous value. Constructing a regression model is all about identifying the relationship between the class label and the input predictors. Multiple problems that can be addressed with a classifier can also be solved using a regression algorithm by defining the outcome as a continuous rather than a categorical variable. Mathematically, regression aims to approximate a function $f$ given a finite sample of training instances $\{(x_1, y_1), ..., (x_n, y_n)\}$. Unlike in classification, however, the range of $y_i$ is not discrete; it can take any value in $\mathbb{R}$. The approximating function $f$ is also called the regression function. A natural way of evaluating the performance of an approximating function $f$ is the residual sum of squares using the following equation:

$$E = \sum((y_i - f(x_i))^2 \tag{2.1}$$

More detailed information on regression analysis can be found in text books, such as [82]. However, the majority of the work presented in this thesis employs classification and clustering. Therefore, most of the focus for this thesis will be on the two paradigms (i.e, classification and clustering).

The classification problem aims to predict group membership (i.e., labels or classes), for a set of observations. The popularity of classification algorithms can be due to the fact that most of the medical diagnostic problems can be simplified into categorical decisions; for example, should a medical patient be treated with medication A, B, or C? The most straight forward

application of the classification problem in the clinical domain is diagnostic classification. In this type of problem, a classifier learns to distinguish, for instance, patients with a particular disease from healthy controls. The classification problem can be further divided into two categories: single-label and multi-label classification, as will be discussed in the next subsections. In this thesis, a variety of classification methods from both the single-label and multi-label classification approaches have been investigated.

#### 2.2.2.1. Single-label Classification Problem

Single-label classification, also called the standard or traditional classification, is a supervised machine learning task where the system learns from a set of labeled input examples to correctly predict the class membership of unlabeled samples. The goal of standard classification is to obtain a model that will be able to assign each of the given unlabeled inputs to the corresponding known output. A more formal definition of classification is:

**Definition 2.2.2.1.** *Given a set of data points* $X = \{x_1, x_2, ..., x_n\}$*, each of them associated with a finite set of target classes* $Y = \{y_1, y_2, ..., y_m\}$*, the classification problem is the task of generating a mapping function* $f : X \rightarrow Y$*, which maps element of* $X$ *to* $Y$*.*

Each training data point $x_i$ is often referred to as instance or sample and is characterized by a finite set of features $F = \{f_1, f_2, \ldots, f_j\}$ that can be either numerical or categorical. Commonly, the features are called variables or attributes. Standard classification concerns with learning from a set of data points that are associated with only one target label from a set of disjoint labels $Y$,with $|Y| \geq 1$. If $|Y| = 2$, then the learning problem is called a binary classification, while if $|Y| > 2$, then it is called a multi-class classification problem [303]. The most common applications of binary classification problems include disease diagnosis, spam, and malware detection, quality control, etc. [84]. Several real-world problems, however, involve the classification of more than two classes. Examples of such problems include the distinction of multiple types of tumor, image classification (e.g., an animal image can be classified as either cat, dog, fox, or rabbit, etc.), character recognition [85] (e.g., classifying an image of a handwritten number into a digit from 0 to 9), biometric identification [86] and face recognition. In general, solving a multi-class problem is more complex and expensive than a binary problem with the same amount of data [87], as the generated model

must be able to separate the given examples into a higher number of categories, which increases the chances of classification errors. As a result, its complexity grows for large number of classes.

There are two ways to solve multi-class classification: using either direct multi-class learning algorithms or decomposition-based approaches that combine several binary classifiers. The former category includes ML algorithms, such as k-nearest neighbors [88], neural networks [89], decision trees [90], and naïve Bayes classifiers [91]. However, some of the well-known ML algorithms, such as support vector machines, logistic regression, and perceptron learning, are originally designed for the solution of binary classification problems. These methods may not be used for multi-class classification directly. In such cases, we used the alternative of decomposing the multi-class problem into binary subproblems. Decomposition is one of the most commonly used approaches to deal with multi-class classification, splitting multi-class problems into a set of binary classification problems. The outputs of binary classifiers are then combined to gain the multi-class outputs.

The most common decomposition strategies used for multi-class classification problems are one-against-all and one-against-one methods [92]. In the one-against-all (OAA) strategy, a multi-class problem with k classes is split into k binary classification problems. Then, k binary classifiers can be created where each classifier is trained to distinguish one class from the remaining k-1 classes. For this approach, the $k^{th}$ classifier is assumed be trained with positive examples belonging to class k and negative examples belonging to the k-1 classes. In one-against-one (OAO) approach, k(k-1)/2 binary classifiers, one for every possible pair of classes, are generated for the given k number of classes. Each example is then classified according to a majority vote amongst the classifiers.

Classification can be a two-step process consisting of model construction (learning) phase and model usage (classification) phase. In the first phase, the training data set is used to train the model where the generated model can be represented as classification rules, mathematical formulas, or decision trees. The second phase refers to using the model for classifying future or previously unknown data points. Sometimes, it can also be a three-step process where a model adjustment step is added between model construction and model usage to adapt the model better to the data. If the model has not been able to classify or predict the concept of the new data, there will be a large generalization or prediction error, which means that the future unseen

samples are not correctly classified. However, when building a classification model, future samples are not accessible; therefore, it is essential to simulate the model using the second phase (i.e., usage phase). Simulating future behavior of the learning model helps to know whether the learning part was successful and to estimate the future generalization error rate. The evaluation of the model through such simulation can be done by dividing the available data into three non-overlapping sets: training, validation, and test sets. First, the model is created by learning from the training set; then, the validation set is used to fine-tune the hyper model parameters. Once the model is trained using train and validation tests, the test set provides an evaluation of the final model. The workflow applied in model usage simulation is illustrated in Figure 2.3. Model validation consists of evaluating how well the classes of the test instances can be predicted.

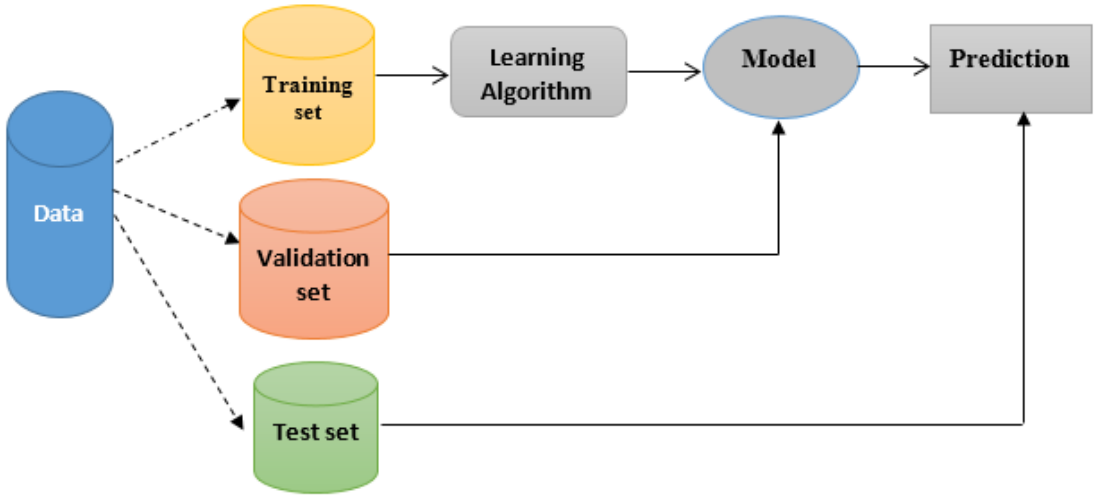


Figure 2.3: A general approach to build and validate a predictive model

For binary classification where there are only two classes (presence and absence), the performance of a model is often evaluated using a confusion matrix (contingency table). In the healthcare domain, the positive class often represents individuals with a medical condition, while the negatives represent healthy individuals or controls. Table 1.1 shows a confusion matrix that summarizes the correct and incorrect predictions for each class. The choice of evaluation metrics affects the objectivity and fairness of the final model assessment. Commonly, most classifiers have been assessed by the overall accuracy rate. However, accuracy alone doesn't reflect false-positive samples, and therefore, it is not particularly appropriate for class

imbalance learning. As a result, some other effective evaluation metrics have been used to assess the classification performance. In a binary classification problem, the outcome of classification performance can be represented by a confusion matrix, as shown in Table 2.1. Then the following performance metrics are obtained from the confusion matrix.

Table 2.1: Outcomes of a binary classification problem

| | | Predicted class | |
|---|---|---|---|
| | | Positive | Negative |
| Actual class | Positive | True Positive (TP) | False Negative (FN |
| | Negative | False Positive (FP) | True Negative (TN) |

**Accuracy** is calculated as the number of all correct predictions divided by the total number of records in the dataset. With accuracy, performance values range between 0 and 1, where 0 is very poor overall accuracy, and 1 is perfect classification accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{2.2}$$

**Sensitivity (Recall or True positive rate)** is calculated as the number of correct positive predictions divided by the total number of (actual) positives. It is the ability of the model to detect the disease if it is really present. The best sensitivity is 1.0, whereas the worst is 0.0.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.3}$$

**Precision (Positive predictive value)** is measured as the number of correct positive predictions divided by the total number of positive predictions. It tells you how often a positive test represents a true positive. The best precision is 1.0, whereas the worst is 0.0.

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

**Specificity** measures the proportion of negatives that are recognized as such. It is calculated as the number of correct negative predictions divided by the total number of (actual) negatives.

$$Specificity = \frac{TN}{TN + FP} \tag{2.5}$$

**G-mean** is the geometric mean of sensitivity and specificity. It is often used when the performance of both classes is expected to be considered.

$$G - mean = \sqrt{Sensitivity * Specificity} \tag{2.6}$$

**F1-score** measures the harmonic mean of recall and precision. It is a metric that takes into account both recall and precision.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.7}$$

In the medical context, model performance is often measured through the analysis of the Receiver Operating Characteristics (ROC) curve. ROC curve is also an important and robust evaluation metrics in the presence of the class imbalance problem. It is a simple graphical representation to evaluate a machine learning model, where it illustrates the performance of a classifier based on the true positive rate (sensitivity) and false-positive rate (1-specificity) [93]. The ROC is a probability curve plotted by varying the threshold set at which the samples are assigned to a specific class. The ROC curve can be summarized into a single value by calculating the area under it, called Area Under the ROC curve (AUC). AUC can be used to assess the performance of predictive models at different threshold settings. It has been known to be the most powerful metric in experimental comparisons of multiple learning algorithms. For binary classification, a perfect classifier should have an AUC of 1, while a random classifier produces an AUC of 0.5 (i.e., the classifier assigns with a 50% chance on one of the two class labels).

#### 2.2.2.2. Multi-label Classification Problem

In many real-world applications, an object may be associated with multiple labels concurrently, and such problems are recognized as multi-label learning[94]. Multi-label classification (MLC)[95] is the task of a multi-label supervised learning paradigm where there is typically a finite set of potential labels that can be applied to the instances of multi-label data (MLD). The basic goal is to simultaneously predict a vector of outputs for a given single input, which means that it is possible to solve more complex decision-making problems. This is opposed to the single-label classification, where each instance is associated with only one label. For a multi-label task, generally, an instance can be associated with a set of labels (i.e., each distinct combination of labels); we call these labels as relevant labels (active labels), while those that are not associated are known as the irrelevant labels. Both

the relevant and irrelevant labels are represented as a binary vector, with a size equal to the total number of labels in the MLD.

MLC is currently receiving considerable attention and is applicable to a variety of research domains, including bioinformatics, text classification, music categorization, medical diagnosis, image, and video annotation [96]. For example, in medical diagnosis, a patient can have multiple side effects for a disease or a medical diagnosis might find a patient suffers from more than one disease at the same time. An image can belong to more than one label; Figure 2.4 shows one of the classical applications of MLC, image labeling. There are four labels in the dataset, and each image can be assigned to any of the labels, or even all of them at once if there was an image in which the four concepts appear. More than one label is assigned to each picture, depending on the elements it contains.

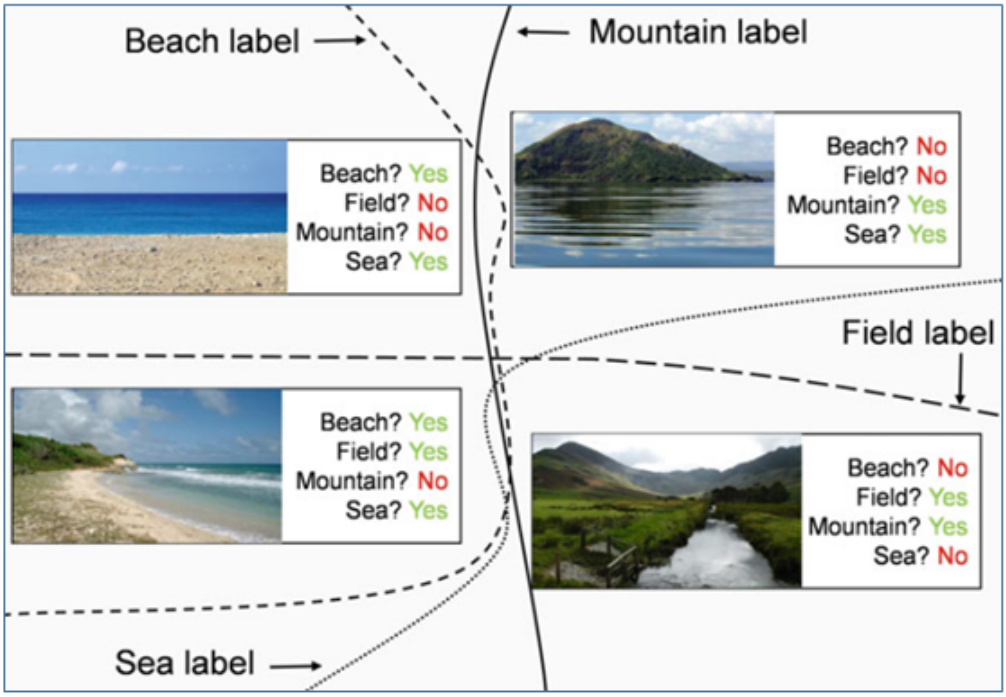


Figure 2.4: Image labeling is an example of a multi-label classification problem [95]

A more complex form of multi-label learning problem closely related to MLC is multi-label ranking (MLR) [97], where the goal is to not only predict a vector of outputs from a finite set of predefined labels but also to rank them according to their relevance to the given input. In a multi-label learning problem, in addition to predicting which labels are relevant and which are

irrelevant, it is often required to get a good ranking of relevant labels (i.e., a list of preferences) from the list of possible labels for each unseen example. MLR is an interesting problem as it subsumes many supervised learning tasks such as multi-label, multi-class and hierarchical classifications [98]. An example application of MLR is document classification, where categories are topics (e.g., technology, politics, and sports) within a document collection (e.g., news article). It is very likely that a document may belong to several topics, and the aim of the learning algorithm is to rank (order) the relevant topics higher than non-relevant ones for a given document query. According to [99], a multi-label problem is assumed to have the following settings:

1. The set of labels is predefined, meaningful and human-interpretable
2. The number of possible labels is limited in scope, and not greater than the number of attributes
3. Each training example is associated with multiple labels of the label-set.
4. The number of attributes may be large, but attribute-reduction strategies can be employed in these cases. The number of examples can also be very large.
5. Labels may be correlated: the relationship between labels represent additional knowledge that can be explored during the training of the learners to facilitate the learning process.
6. Dataset can be unbalanced: this can be viewed from three perspectives: imbalance within labels, imbalance among labels, and imbalance among label-sets.

#### 2.2.2.2.1. Formal Definition of Multi-label Learning

The formal definition of multi-label learning is presented as follows [249]: Let $X$ be a d-dimentional input space of categorical or numerical features and an output space of $q$ labels $L = \{\lambda_1, \lambda_2, ..., \lambda_q\}$,$q > 1$. A multi-label example can be defined as a pair $(x, Y)$ where $x = (x_1, x_2, \ldots, x_q) \in X$ and $Y \subseteq L$ is called a label-set. $D = \{(x_i, Y_i)|1 \leq i \leq m\}$ is a multi-label dataset (MLD) composed of a set of $m$ instances. Let $Q$ be a quality criterion which rewards models with high predictive performance and low complexity. If the task is an MLC, then the goal is to find a function $h : X \rightarrow 2^L$ such that $h$ maximizes $Q$. If the task is an MLR, then the goal is to find a function $f : X$ x $L \rightarrow \mathbb{R}$ such that $f$ maximizes $Q$, where $\mathbb{R}$ is the ranking of labels

for a given example. Table 2.2 shows an example of a single-label dataset and an MLD.

#### 2.2.2.2.2. Multi-label Learning Methods

The different methods proposed in the literature to deal with multi-label learning problems [96] can be categorized into three groups: problem transformation, algorithm adaptation, and ensemble methods. Figure 2.5 shows the different categories and associated methods.

### Problem Transformation

Problem transformation is one of the simpler strategies which converts the multi-label problem into one or more single-label (i.e., multi-class or binary) problems. It is the idea of preprocessing an MLD to generate a dataset that can be used by any off-the-shelf binary or multi-class classifiers. Often, the outputs produced by those classifiers have to be back-transformed into the subsets of labels to obtain the multi-label prediction. The basic approaches for the problem transformation method can be grouped into three categories: binary relevance, label powerset, and pairwise methods.

Table 2.2: Example of labels in a single-label and multi-label datasets

| Sample | Features | Single-label output | | Multi-label output | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Binary | Multi-class | | | | | |
| | | $Y \in L = \{0,1\}$ | $Y \in L = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $Y \subseteq L = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ |
| 1 | $\overline{X_1}$ | 1 | $\lambda_1$ | 1 | 1 | 0 | 1 | $\{\lambda_1, \lambda_2, \lambda_4\}$ |
| 2 | $\overline{X_2}$ | 0 | $\lambda_2$ | 0 | 0 | 0 | 1 | $\{\lambda_4\}$ |
| 3 | $\overline{X_3}$ | 0 | $\lambda_3$ | 0 | 1 | 1 | 1 | $\{\lambda_2, \lambda_3, \lambda_4\}$ |
| 4 | $\overline{X_4}$ | 1 | $\lambda_4$ | 1 | 0 | 1 | 0 | $\{\lambda_1, \lambda_3\}$ |
| 5 | $\overline{X_5}$ | 0 | $\lambda_5$ | 0 | 1 | 1 | 0 | $\{\lambda_2, \lambda_3\}$ |

***Binary Relevance (BR)***: is the baseline approach that decomposes the multi-label problem into $q$ binary independent problems by learning one classifier for each label, using all the instances associated with that label as positive and all the remaining samples as negative. When making a prediction, each binary classifier predicts whether its label is relevant for the given example or not, resulting in a set of relevant labels. The final multi-label prediction for a new instance is determined by aggregating the classification

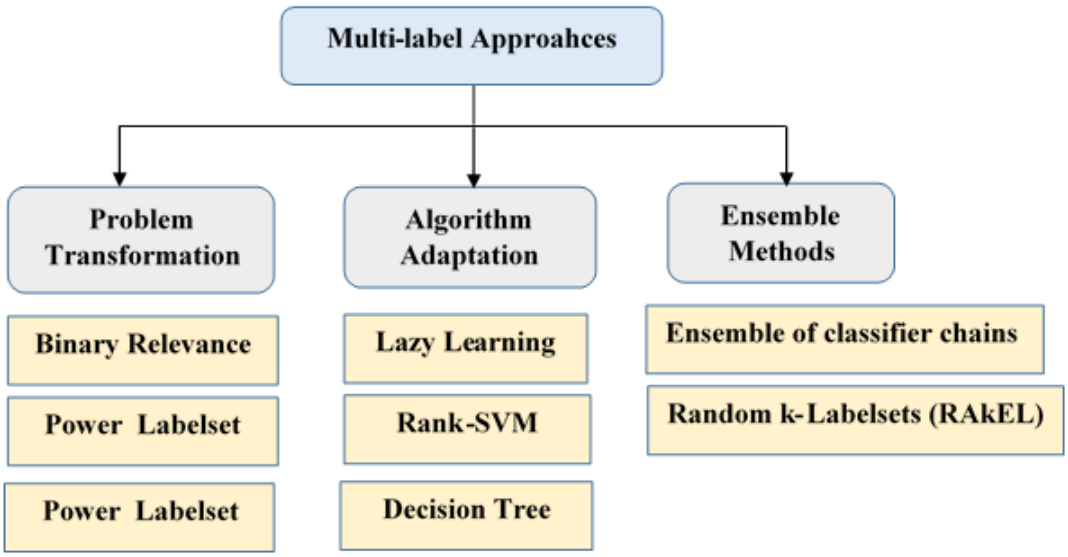


Figure 2.5: Multi-label Learning Approaches

results from all the independent classifiers. During the ranking task, the labels are ranked according to the probability associated with each label by the respective binary classifier. The BR method is an extended form of the one-against-all (OAA) approach, which has been used for facing multi-class classification by means of an ensemble of binary classifiers. Although BR is relatively simple to implement, it is realized that BR ignores the possible relationship between labels (such as label dependency, co-occurrence, and correlation). To deal with the limitation of the BR method, the classifier chain (CC) was introduced in [101] involves $q$ binary classifiers linked along the chain , which then resolves the disadvantage of BR method by taking into account the relationship among labels.

***Label Powerset (LP)***: is the most natural approach of problem transformation method, which combines the entire label-sets into single (atomic) labels to form a single-class classification problem. In the LP method, the interrelationships among labels are mapped directly from the data, and therefore consider label correlations, as all the existing combinations of single-labels present in the training instances are used as a possible label in the correspondent multi-class classification. The complexity of LP depends on the complexity of the single-label classifier with respect to the number of classes. Although it takes into account the label correlations, it suffers from

the increasing complexity that depends on the number of distinct label-sets. The large number of distinct label-sets can also lead to the problem of imbalanced label-sets. In order to resolve this problem, pruned problem transformation (PPT) has been developed by Read [102], which chooses only the transformed labels that occur more than a predefined number of times. PPT tries to solve LP's problems related to unbalanced data and complexity, by pruning samples with less frequent label-sets by focusing on the most important label-sets through the use of a user-defined threshold. Another LP-based approach is HOMER [103], which first constructs a hierarchy of the multiple labels and then builds a classifier for the label-sets in each node of the hierarchy.

***Pairwise methods***: the third problem transformation method to solve multi-label learning is the ranking by pairwise comparison (RPC). This approach is similar to the one-against-one (OAO) approach for the multi-class classification problem. The basic idea is to transform the multi-label dataset into $q(q-1)/2$ binary datasets, one for each pair of labels, and a binary classifier is built for each dataset. Given a new instance, all models are invoked, and a ranking is obtained by counting votes for each label. The main disadvantage of RPC is the space complexity and the need to query all the generated binary models at run time. This quadratic complexity makes RPC very sensitive to a large number of labels and often intractable for large problems. Calibrated label ranking (CLR) extends the RPC by introducing a virtual label which indicates the boundary between the relevant and irrelevant labels. The final ranking includes the virtual label that acts as a split point for the relevant and irrelevant labels obtaining a consistent ranking and bipartition.

### Algorithm adaptation

Algorithm adaptation methods adapt, extend, and customize existing single-label learning techniques to handle the task of multi-label learning directly. The adaptations of several single-label learning algorithms have been proposed in the literature based on the following machine learning algorithms: decision tree, boosting, k-nearest neighbors, neural networks, and support vector machines. The adapted methods can be able to handle the multi-label data directly.

***Decision trees***: Some of the decision tree algorithms have been adapted to multi-label classification problems. In [104], the C4.5 algorithm is extended for multi-label data (ML-C4.5) by modifying the formula of entropy. A

large number of leaves are generated for all combinations of different labels. This method can handle multiple labels on several levels of the hierarchy and assign a higher cost to misclassifications higher up in the hierarchy.

***Boosting***: AdaBoost.MR and AdaBoost.MH [105] are extensions of the well-known AdaBoost algorithm for multi-label learning. AdaBoost.MR is designed specifically to find a hypothesis which ranks the correct labels at the top of the ranking, while AdaBoost.MH is designed to minimize Hamming loss. Besides, AdaBoost.MH can be combined with the alternating decision trees algorithm to produce the Adapted Decision Tree Boosting [106]. The resulting multi-label models of this combination can be interpreted by humans.

***K-Nearest Neighbors***: the popular k-Nearest Neighbors (kNN) lazy learning algorithm has been adapted to multi-label learning problems. In [175], multi-label k-nearest neighbors (MLkNN) is proposed using a Bayesian approach. It uses the Maximum A Posteriori Principle (MAP) to determine the relevant label-set for a new given instance, based on prior and posterior probabilities for the frequency of each label within the k nearest neighbors. It also has the capability to produce the ranking of the labels.

***Neural networks***: artificial neural networks have also been extended for the multi-label learning task. BPMLL [108] is an extension of the popular back-propagation algorithm for multi-label learning. The main modification to the algorithm is the introduction of a new error function that takes multiple labels into account.

***Support vector machines***: authors in [109] have proposed a ranking approach, called Rank-SVM, for multi-label learning that is based on SVMs. It incorporates pairwise label constraints directly in the optimization problem, which is used to minimize the ranking-loss. The main function they use is the average fraction of incorrectly ordered pairs of labels.

### Ensemble Methods

Ensemble methods are developed on top of problem transformation or algorithm adaptation methods. The most common problem transformation ensembles are the ensemble of classifier chains (ECC) [110], RAkEL (random k-label-sets) [245], and an ensemble of pruned sets [112]. The combined algorithms can be homogeneous, i.e., the ensemble can be built from a collection of similar classifiers, or heterogeneous when various classifiers contribute to building the ensemble. The ensemble approach may reduce the limitation of one classifier by adding an ensemble of classifiers.

***ECC*** is an ensemble approach that uses a combination of classifier chains (CC) to produce the ensemble model. ECC was proposed to alleviate the effect of classifier order in CC, by training an ensemble of CC classifiers. Each CC can be trained with a random chain ordering on a random subset of training patterns.The final prediction is gained by aggregating the predictions by the label and then using a threshold for selecting the relevant labels. Both CC and ECC pass label information between binary classifiers, and they take into account correlations among labels and overcome the limitation of BR, which ignores such correlations. Furthermore, ECC reduces the risk of selecting a bad chain ordering, which can lead to a bad prediction performance of the classifier. The diversity in ECC is produced by using various chains and by selecting random subsets of instances.
***RAkEL*** constructs an ensemble of classifiers using LP as the base classifier. It splits the large label-sets into several smaller models or subsets, which are related to random and small-sized k-label-sets. Given a new instance, the results of all LP classifiers are merged by applying a majority voting process for each label to determine the final set of labels. RAkEL has several advantages over LP. First, the LP tasks of each classifier are much simpler since they only consider a small subset of the labels. Also, the base classifiers include a much more balanced distribution of classes than using LP with the full set of labels. Further, RAkEL allows predicting a label set that does not appear in the original training set. A variation of RAkEL, called RAkEL++ [113], uses the confidence values of each classifier instead of bipartitions in order to generate the final prediction for each label. Another variation, called RAkELd [245], generates disjoint subsets of k labels, taking into account each label exactly once and reducing the complexity of other RAkEL variants. The diversity in all the variants of RAkEL is generated by a different selection of labels in each classifier.

### 2.2.3 Unsupervised Learning

Unlike supervised learning tasks, unsupervised machine learning works directly on an unlabeled dataset. The goal is to discover the underlying structures in the data. In the absence of labels to guide the training process, the labels can be uncovered by the learning algorithm. The next paragraphs will describe the most common unsupervised methods employed in a clinical context.

#### 2.2.3.1. Cluster Analysis

Cluster Analysis is the fundamental task in unsupervised learning, where the input is a set of samples, each described by a vector of attribute values (but no class labels). The output is a set of two or more clusters of samples. Clustering is the process of partitioning a set of observations into distinct groups so that the observations within each group (i.e., clusters) are quite similar to each other, while observations in different groups are quite different from each other. Clustering is probably the most used exploratory data analysis method across different domains and is often applied to get an intuition about the underlying structure of the data, for finding meaningful groups, and also for feature extraction and summarizing. More formally, given a data space $X$, clustering can be thought of as a partitioning of this space into $K$ distinct parts, i.e.,$f : X \to \{1, \ldots, K\}$. This partitioning is done by optimizing some internal clustering criteria such as the intracluster distances, etc. The value of $K$ is usually found by employing a second criterion that measures the robustness of the partitioning.

#### 2.2.3.1.1. Procedures of Clustering Analysis

The clustering procedure may result in different partitioning of a given dataset, depending on the specific criteria used for clustering analysis. Thus, there is a need for some preprocessing before performing a clustering task in a dataset. A typical clustering process is presented in Figure 2.6 and can be summarized in the following steps [114]. These steps are closely related to each other and affect the generated clusters.

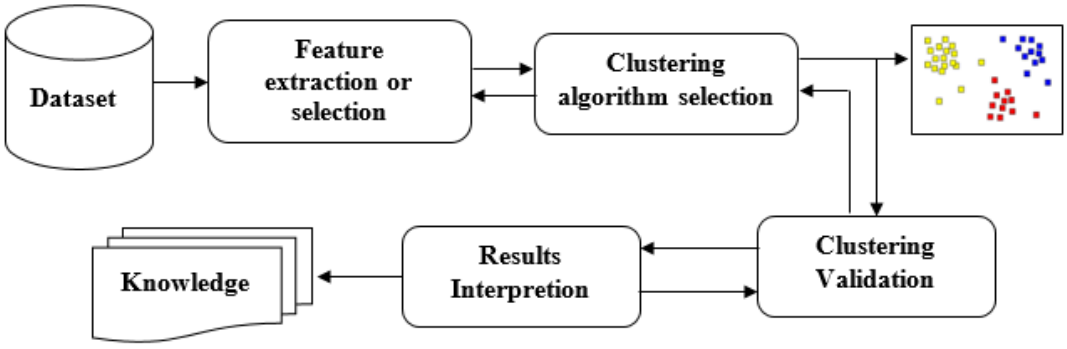


Figure 2.6: A typical cluster analysis consists of four steps

***1. Pattern representation (Feature extraction or selection)***

Pattern representation refers to the number of available patterns and the

type and scale of the features available to the clustering algorithm. Feature extraction is the process of using one or more transformations of the input features to generate new principal features. Feature extraction can be elaborated in the context of dimensionality reduction and data visualization [115]. Feature selection is the process of finding the most effective subset of the original features to use in clustering. Either or both of these methods can be used to obtain an appropriate set of features to use in clustering.

### *2. Clustering algorithm design or selection*

The clustering step is usually combined with the selection of a corresponding proximity(i.e, the closeness or distance) measure and the construction of a clustering criterion function (i.e, finding the optimal partitioning of a data set according to some criterion function or agorithm).
***i.Proximity measures***: the definition of pattern proximity measure appropriate to the data domain is an important step, as it directly affects the formation of the resulting clusters. Almost all clustering techniques are implicitly or explicitly connected to some definition of proximity measure.
***ii.Clustering criterion***: once a proximity measure is chosen, the construction of a clustering criterion function makes the partition of clusters as an optimization problem. Clustering is everywhere, and plenty of clustering algorithms have been developed to resolve different problems in specific fields. However, there is no universally accepted clustering algorithm that can be used to solve all types of problems. As it is proved through an impossibility theorem, "it has been very difficult to develop a unified framework for reasoning about clustering at a technical level, and profoundly diverse approaches to clustering" [116]. Therefore, it is essential to carefully explore the characteristics of the problem at hand in order to select or design an appropriate clustering method.

### *3.Cluster validation (Assessment of results)*

Given a dataset, any clustering algorithm can usually generate clusters, no matter whether the structure exists or not. Also, different methods often lead to different clusters; and even for the same algorithm, parameter identification, or the presentation order of input patterns may affect the final results. Thus, effective evaluation standards and criteria are essential to provide the users with a degree of confidence for the clustering results derived from the used algorithms. These evaluation methods should be objective and have no preference for any algorithm. Moreover, they should

be beneficial to answer questions, such as how many clusters are hidden in the dataset, what is the reason for choosing a specific algorithm instead of another, or if the clusters generated are meaningful or just an artifact of the algorithms. In general, there are three categories of criteria for testing the goodness of clustering results: external criteria, internal criteria, and relative criteria. These are defined on two types of clustering structures, called partitioning clustering and hierarchical clustering [117]. External criteria (i.e., indices) are based on some prespecified structure, which compares the clustering result to a reference result, that can also be considered as the ground truth. If the result is somehow similar to the reference, we regard the final output as a "good" clustering. However, the reference result is not provided in most real applications. Internal indexes are not dependent on external information where the evaluation of clustering is compared only with the result itself, i.e., the structure of the generated clusters and their relations to each other. Unlike the external validation, internal clustering validation is only based on the inherent information of the data, which measures the compactness and the separation of clusters. Compactness measures how closely data points are grouped in a cluster. Compactness is often based on distances between in-cluster points. The most popular way of calculating the compactness is through variance, i.e., the average distance to the mean, to estimate how objects are bonded together with its mean as its center. A small variance indicates high compactness (Figure 2.7). Separation measures how different are the obtained clusters from each other. Users of clustering algorithms are interested in the well-separated clustering results (Figure 2.8). The third method of clustering evaluation is based on relative criteria. The basic idea is the validity of a clustering structure by comparing it to other clustering schemes, resulting from the same algorithm but with different parameter values. An in-depth survey on clustering validity can be found in [118]. Generally, the aim of clustering is to identify the intrinsic divisions in a dataset, and the validation criteria provide some insights on the quality of clustering results, yet how to determine what comprises a good clustering has still a problem requiring more effort. It can be shown that there is no absolute 'best' criterion, which could be independent of the final goal of the clustering. Thus, it is the user who must provide this criterion in such a way that the result of the clustering will suit their needs [119].

### *4.Results Interpretation*

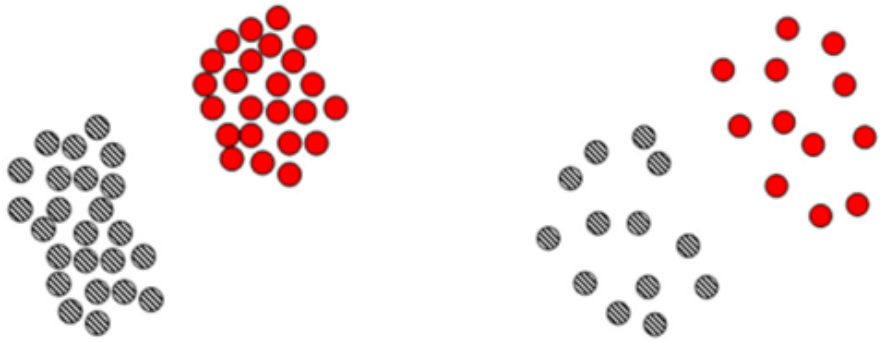

Figure 2.7: Clusters on the left have better compactness than the ones on the right

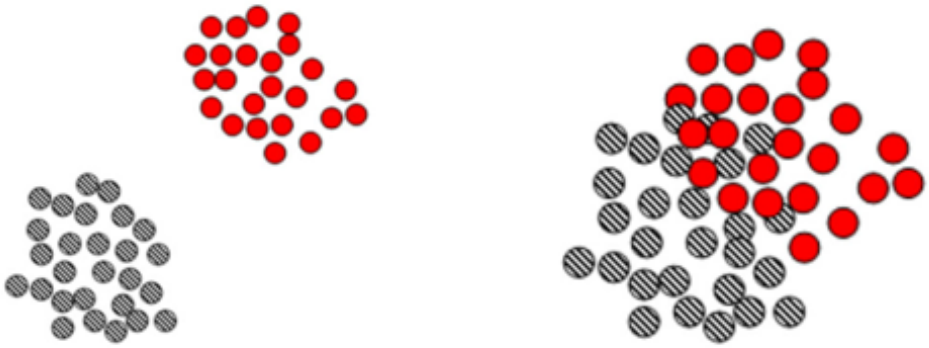

Figure 2.8: Clusters on the left have better separation than the ones on the right

The final target of clustering is to supply users with meaningful perceptions from the original dataset, with the aim that they can effectively solve the problems faced. Experts in different domains interpret the data groupings. Further analyses, even experiments, may be required to assure the reliability of extracted knowledge.

#### 2.2.3.1.2. Categorization of Clustering Algorithms

Different initialization parameters and varying criteria settings often lead to different taxonomies of clustering algorithms. Various approaches to clustering data can be described broadly with the help of the hierarchy shown in Figure 2.9. Other categorizations of clustering algorithms are also possible; ours is based on the discussion in [120].

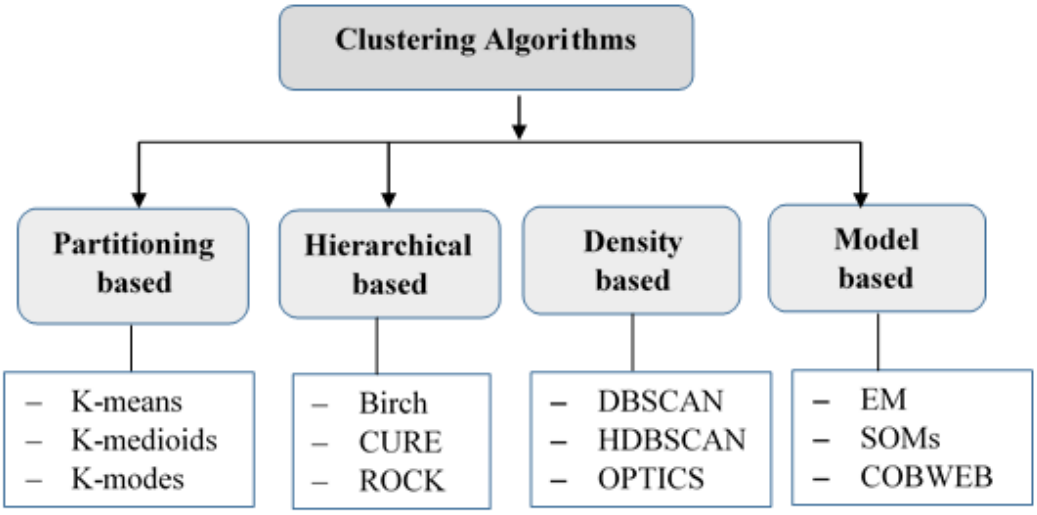


Figure 2.9: An overview of clustering taxonomy

**Partitioning-based**: these are simply divisions of the set of data points into non-overlapping clusters such that each object is in exactly one group. In other words, the partitioning methods split data objects into a number of groups, where each group constitutes a cluster. These clusters must fulfill the following requirements: (1) each partition must comprise at least one data point, and (2) each object must belong to exactly one group. The most common and widely used partitioning methods are k-means, k-medoids and K-modes, and their variations. The k-means algorithm takes the input parameter, k, and divides a set of n objects into k clusters so that the resulting inter-cluster similarity is low, while the intra-cluster similarity is high. Cluster similarity is measured considering the mean value of the data points in a cluster, which can be viewed as the cluster's centroid or center of gravity. In the K-medoids algorithm, objects which are near the center represent the clusters. It is the most centrally located object of the cluster, with a minimum sum of distances to other points. K-modes method extends the k-means paradigm to cluster categorical data by replacing the means of clusters with modes, using new dissimilarity measures to deal with categorical objects and a frequency-based method to update modes of clusters. The k-means and the k-modes methods can be integrated to cluster data with mixed numeric and categorical values. An interested reader can get an in-depth discussion of partitioning methods in [121].

**Hierarchical-based**: data are organized in a hierarchical manner depending on the medium of proximity. Proximities are obtained by the intermediate nodes. A dendrogram represents the datasets, where individual data is presented by leaf nodes. The initial cluster gradually partitions into many clusters as the hierarchy continues. Hierarchical clustering approaches can be agglomerative or divisive. An agglomerative clustering begins with one object for each cluster and recursively merges two or more of the most appropriate clusters. A divisive clustering starts with the dataset as one cluster and recursively splits the most appropriate cluster. The process continues until a stopping criterion is reached (frequently, the requested number k of clusters). The hierarchical approaches have a major drawback, though, which relates to the fact that once a step (merge or split) is performed, this cannot be undone. CURE, BIRCH, ROCK, and Chameleon are some of the well-known algorithms of this category [122].

**Density-based**: here, data points are divided based on their regions of density, connectivity, and boundary. They are closely related to point-nearest neighbors. A cluster, defined as a connected dense component, grows in any direction that density leads to. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes. Also, this provides natural protection against outliers. Thus, the overall density of a point is analyzed to determine the functions of datasets that influence a particular data point. According to Hartigan [123], in density-based clustering, clusters are the high-density regions separated by contiguous regions of a low density of objects. DBSCAN, HDBSCAN, OPTICS, DBCLASD, and DENCLUE are algorithms that use such a technique to filter out outliers (noise) and discover clusters of arbitrary shape [124]. Among these, DBSCAN and HDBSCAN have been tried in this thesis in order to discover plausible clusters that may exist in the elderly data. The difference between these two methods is that DBSCAN can only provide a flat (non-hierarchical) clustering of the data objects, based on a global density threshold, while HDBSCAN is a hierarchical extension of DBSCAN algorithm for varying epsilon threshold values [125]. Unlike DBSCAN, HDBSCAN allows finding clusters of variable densities without using user-defined distance threshold epsilon. However, there are cases where an epsilon threshold can be advantageous.

**Model-based**: such a method optimizes the fit between the given data and

some (predefined) mathematical model. It is based on the assumption that the data is generated by a mixture of underlying probability distributions. Also, it leads to a way of automatically determining the number of clusters based on standard statistics, taking noise (outliers) into account, and thus yielding a robust clustering method. There are two major approaches that are based on the model-based method: statistical and neural network approaches. MCLUST is probably the best-known model-based algorithm, but there are other good algorithms, such as EM (which uses a mixture density model), conceptual clustering (such as COBWEB), and neural network approaches (such as self-organizing feature maps). The statistical approach uses probability measures in determining the concepts or clusters. Probabilistic descriptions are typically used to represent each derived concept. The neural network approach uses a set of connected input/output units, where each connection has a weight associated with it. Neural networks have several properties that make them popular for clustering. First, neural networks are inherently parallel and distributed processing architectures. Second, neural networks learn by adjusting their interconnection weights so as to best fit the data. This allows them to normalize or prototype [126].

Clusters can also be categorized based on other dimensions, such as the type of clustering, along with which one can describe clustering paradigms and methods. Based on type, clusters can be classified as hard clustering where each object either belongs to a cluster completely or not, and soft clustering where instead of placing each object into a separate cluster, a probability or likelihood of that object to be in those clusters is assigned [127]. In healthcare, clustering represents an essential tool; it has been used for solving several different problems such as: create a taxonomy of living things, identify clusters of genes with similar biological functions, stratify patients with similar clinical characteristics, etc.

#### 2.2.3.2. Dimensionality Reduction

Dimensionality reduction is the preprocessing procedure to remove redundant features, irrelevant and noisy data, in order to improve learning feature accuracy and reduce training time. Dimensionality reduction can be implemented using feature selection or feature extraction. Feature selection is based on selecting those features which contribute most to class separability. In other words, redundant and irrelevant features are ignored. Feature extraction considers all the information content and maps useful information content into a lower-dimensional feature space. Many non-linear

dimensionality reduction techniques have recently been developed to improve upon linear techniques like PCA in feature extraction from complex non-linear data manifolds. The following are some of the commonly used dimensionality reduction techniques in the machine learning domain.

**Principal Component Analysis (PCA)** is a linear distance preservation technique that aims to summarize the original set of features into a smaller set that collectively explains most of the variability in the original variables. PCA uses an orthogonal transformation to convert a set of data points, of possibly correlated variables, in a set of values of linearly uncorrelated variables called principal components. Thus, the total number of components is less or equal than the number of original features. PCA is one of the oldest and best-known methods in multivariate analysis and data mining [128].

**Multiple Correspondence Analysis (MCA)** [129] is an extension of the simple correspondence analysis for summarizing and visualizing a data table containing more than two categorical variables. It can also be viewed as a generalization of PCA when the variables to be analyzed are categorical instead of quantitative.

**t-Distributed Stochastic Neighbour Embedding (t-SNE)** is a non-linear dimensionality reduction method for visualizing high-dimensional data which converts similarity between data points to joint probabilities and preserves the local data manifold. It is a special case of stochastic neighbor embedding but uses a different cost function to make its application to larger datasets practical [130].

In dimensionality reduction techniques, like PCA, there are four criteria that may be used to determine the number of meaningful components to retain: the eigenvalue-one criterion, the scree test, the proportion of variance accounted for, and the interpretability criterion [131].

***The eigenvalue-one criterion***: in PCA, one of the most widely used criteria for solving the number of components problem is the eigenvalue-one criterion. This approach helps to retain and interpret any component with an eigenvalue greater than 1.0. Each observed feature contributes one unit of variance to the total variance in the dataset. Any component that shows an eigenvalue greater than 1.0 is accounted for a higher amount of variance than has been contributed by one variable. Such a component is, therefore, accounting for a meaningful amount of variance, and is worthy of being retained.

***The scree test***: the scree test allows us to plot the eigenvalues linked to

each component and consider a break between the components with fairly greater eigenvalues and those with smaller eigenvalues. The components that are appeared before the break are assumed to be meaningful and are kept for rotation; those appearing after the break are assumed to be unimportant and are not retained. An interested reader can get more information in [131].

***The proportion of variance accounted for***: a third criterion in determining the number of factors involves retaining a component if it accounts for a specified proportion (or percentage) of variance in the dataset. For example, one may choose to retain any component that accounts for at least 10% of the total variance. more detailed information can be obtained in [131].

***The interpretability criteria***: probably, the essential criterion for selecting the number of components is the interpretability criterion: interpreting the significant meaning of the retained components and verifying that this interpretation makes sense in terms of what is known about the constructs under investigation. Some detailed procedures on using the interpretable criteria can be found in [131].

#### 2.2.3.3. Association Rule Mining

Association rule mining aims to find frequent and interesting patterns or associations among the observations in a dataset. Association rules are defined as an implication of the form:

$$X \rightarrow Y$$

where $X, Y \subseteq I, I = i_1, i_2, ..., i_n$ is a set of attributes called item. A set of items is referred to as an itemset. A transaction database, T, is a finite set of transactions $T = \{t_1, t_2, \ldots, t_n\}$, where each transaction contains a set of items. The goal is to determine frequent patterns, associations, correlations, or causal structures contained in the item sets in the transaction database and express these relationships in terms of association rules, if-then rules.

Association rules do not differ much from the classification rules, except that they predict any attribute, not simply the class attribute, including a combination of them. The mining of frequent itemsets leads to the discovery of associations and correlations among items in large relational or transactional datasets. A famous association rule, which emerged from the analysis of supermarket shoppers, is the ***market basket analysis***. This process analyzes customer buying habits by discovering associations between the

various items that customers place in their shopping baskets. The finding of such association rules can be helpful for retailers to develop marketing strategies by gaining insights into which items are commonly purchased together by customers. For instance, if customers buy milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space. Apriori is a classic algorithm for learning association rules [132]. The algorithm tries to find subsets of attributes which are in common to at least a minimum number instances. Using a bottom-up approach, frequent subsets are extended one item at a time, and groups of candidates are tested against the data. Apriori terminates when no further extensions can be found. Many other algorithms have been proposed after the Apriori algorithm [133].

## 2.3 Machine Learning in Healthcare

The clinical decision-making process in several aspects of the healthcare system is often complex and requires many considerations before arriving at a course of action inpatient care. Clinical diagnosis and prognosis processes depend on the experience, judgment, emotions, intuitions, and knowledge of physicians, which tends to be highly subjective and varies from person to person. Detecting adverse clinical events by eliminating some degree of physician's subjectivity, reducing mortality rates, mitigating healthcare costs and medical complications at various hospitals, all represent important research challenges in the healthcare domain [134]. Healthcare setting is generally perceived as "data-rich," derived from a wide variety of sources such as electronic records (clinical and administrative data), images, sensors, and text in the form of biomedical literature/clinical notes. This variety in the data collection and representation procedures leads to many challenges in both the processing and analysis of the underlying data. There is also a wide heterogeneity in the methods that are required to analyze these different forms of data. In addition, the diversity of data naturally creates various data integration and data analysis challenges.

Recently, the massive potential of such integrated data analysis approaches are being realized, and rapid advances in technologies have changed the paradigm of healthcare sector [135] where medical experts are using machine learning (ML) algorithms for diagnosis and treatment recommendations, patient engagement and adherence, administrative activities, and prediction

of outcomes in many clinical scenarios. ML plays a significant role in the healthcare domain and is being increasingly applied to develop complex models, and extract medical knowledge, exposing novel ideas to practitioners, and specialists [136]. As there is no generalized ML model, various models have been devised to assist with the different decision-making processes. Some of these learning methods can be used to increase our understanding of the current world (e.g., identifying risk factors for infections or identifying patients who may be for the development of a particular condition), whereas others focus on predictions about the future (e.g., predicting who will become at high risk of hospitalizations or mortality). ML predictive models can highlight enhanced rules in decision-making regarding individual patient care. These are also capable of autonomous diagnosis and treatment of different diseases under clinical regulations. The use of ML models is likely to increase as healthcare providers, and patients seek to improve their clinical decision-making process to achieve better outcomes while reducing overall healthcare costs [137]. However, it should be clear that the final clinical decision should be made by the physicians as humans are more flexible and capable of identifying outlying details that the ML system is unable to account for (e.g., due to the lack of certain information) [138]. Hence, ML models should serve as guidelines aiming to assist the overall standard of healthcare and should not be used as a replacement for physicians. An ideal scenario is to capitalize on the highly accurate predictions that an ML-based system can offer while allowing physicians to have full flexibility and responsibility in making a good clinical decision [139].

The well-known benchmark ML methods that have been employed to healthcare includes Artificial Neural Network (ANN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR)[140], and Genetic Programming (GP)[141]. The use of these algorithms for extracting insights from large medical databases is invaluable as medicine is a domain that is complex and difficult to model by humans. These techniques are capable of handling a large amount of data from different sources, incorporate expert knowledge into the analysis, offer data-driven predictions that can assist clinicians in making their decision. The following are some examples of applications that have adopted ML techniques as an approach to enhance clinical decision making:

1. ANN was applied for disease diagnosis based on patients' history database, comparing predictive accuracy of various types of ANN and statistical models for diagnosis of coronary artery disease [142], diagnosis and risk

group assignment for pulmonary tuberculosis among hospitalized patients [143], and non-invasive diagnosis of early risk in dengue patients [144]. Applications of ANN for prediction includes developing a risk prediction model to predict the chances of diabetes complication according to changes in risk factors [145], identifying the optimal subset of features from a given set of features for diagnosis of heart disease [146], modeling daily patient arrivals in the emergency department. In [147], multiple ANNs approach is proposed to estimate the probability of nosocomial infection. Multiple ANNs was constructed by connecting individual ANNs (that predicts the probability of nosocomial infection at different time period) sequentially, where the output of an ANN is connected to the input of the next ANN. Experimental results show that with multiple ANNs, it outperforms multivariate regression models in predicting the risk of nosocomial infection.

2. Authors in [148] proposed an e-doctor, a web-based application that makes automatic diagnoses about health problems based on the support vector machine (SVM). System administrators define specific characteristics for each medical problem that can be diagnosed, and train the SVM by entering sample files of statistical data. After that, health staff can feed exam information about patients, and e-doctor makes an automatic diagnosis or prediction by means of answering if the patient has (or may have in the future) a specific health problem. The application can be used in cases where statistical information plays a vital role in deciding about a patient's condition. A prototype was developed, and the system trained and tested for the case of heart symptoms. Another study in [149] was employed using SVM with polynomial kernel for the estimation of whether tracheal intubation could be simple or difficult before anesthesia is carried out. A total of 264 medical records and 13 physical variables were analyzed. The use of 13 basic and anthropometrical features has a significant advantage over the approach taken by some anesthetists where a single feature is examined ahead of anesthesia. This is because most specialists agree that full consideration of multiple variables would improve the prediction accuracy of physical airway examination. Based on 4-fold cross-validation, an average classification accuracy of 90.53 % was achieved in the study.

3. Random forest (RF) has been incorporated in a variety of applications such as risk management, tailored health communication, and decision

support systems in healthcare [150]. RF has also been applied in a variety of health conditions, such as for diabetic retinopathy classification and detection [151], breast cancer diagnosis [152], predicting healthcare-associated infections [153], and healthcare monitoring system [154]. It is highly nonlinear and works well with high-dimensional data, clearly outperforming classic statistical methods.

4. Decision Tree (DT) is an important method for decision making and risk analysis, which is often represented in the form of a graph or list of rules. One of the most important features of decision trees is the ease of their application. Decision trees are a reliable and effective decision-making technique that has been used in different areas of medical decision making (classification, diagnosing, etc.). Interesting applications of decision trees have been used in medical and health care domains for more than 20 years [155].

5. Logistic Regression (LR) is an ML method for modeling dichotomous outcomes, and it is one of the most widely used techniques in medical decision making starting from the early 70s to date [156]. LR is also widely applicable to epidemiologic studies concerned with quantifying an association between a study factor (i.e., an exposure variable) and a health outcome (i.e., disease status) [157]. An in-depth survey of LR to clinical applications can be found in [158].

6. Nature-inspired computing algorithms, such as evolutionary algorithms, swarm intelligence, are emerging approaches that are based on the principles and inspiration of the biological evolution of nature to develop new and robust competing techniques [159]. Genetic programming (GP) is an evolutionary algorithm that has been applied to solve various real-world problems. In the medical domain, GP has been used in the diagnosis and monitoring of Parkinson's disease, Alzheimer's disease and other neurodegenerative conditions [160], prediction of cervical spine disease [166], in the early detection of breast cancer through automated assessment of mammograms [162] and much more.

Although the current role of ML-based clinical decision-making system solves about patient diagnosis, prognosis, and image analysis, it is suggested that it has great potential to improve considerable aspects of clinical health care in the future, which includes (1) recommendation

of the most appropriate and cost-efficient diagnostic process, (2) personalization of therapeutic strategies that maximize efficacy and safety, (3) real-time and transparent monitoring of patients' health, and (4) discovery of new medical knowledge that has a direct and profound impact to the quality of patients' health and care [163]. On the other hand, ML does not explain why the result was gained or its underlying mechanism. If unexpected results are generated, the data must be re-investigated to identify whether technical or human errors have created biases, followed by careful interpretation and validation in the context of the problem at hand.

# Chapter 3

# Imbalanced Data Classification: A Systematic Review

This chapter provides a review of the approaches for handling imbalance problem in multi-label classification by collecting the existing research work. As the first systematic study of addressing an imbalanced problem in MLC, this chapter presents a comprehensive survey of the state-of-the-art methods for imbalanced MLC, including the characteristics of imbalanced multi-label datasets, evaluation measures and comparative analysis of the proposed methods. Numerous articles related to imbalanced MLC published between 2006 and 2019 were collected and reviewed. Several methods and techniques that have been proposed to overcome the negative impacts of imbalanced MLC have been reviewed: resampling methods, classifier adaptations, ensemble methods, and cost-sensitive learning approaches. This chapter also discusses important results reported so far in the literature and highlights some of their strengths and limitations to guide future research.

## 3.1 Introduction

Classification is one of the most important machine learning topics [164]. The main goal is to train a computational model using a set of labeled samples, obtaining a model that is able to correctly classify new unlabeled samples. Traditional single-label classification is one of the most well-established and adopted machine learning paradigms. It provides fast and accurate predictions and is successfully applied in many application domains [165, 166].

Binary and multi-class classifications are subcategories of single-label classification that offer learning from a set of samples that are associated with a single label. Unlike traditional classification methods, multi-label classification (MLC) maps a set of relevant labels to an instance simultaneously [303, 168]. Recently, MLC has gained much importance and attracted research attention with a wide range of applications, including medical diagnosis, music categorization, emotion recognition, text classification, and image/video annotation [169, 170]. In all these cases, the task is to assign a label set for each unseen instance. For example, in bioinformatics, one gene sequence can be associated with a set of multiple molecular functions [171]. In-text categorization, a new article can cover multiple aspects of an event, thus being assigned with a set of multiple topics [172].

There are two well-known approaches for solving an MLC task: problem transformation and algorithm adaptation methods. The former transforms the MLC task into one or more single-label classification [173], or label ranking [174] tasks , while the latter aims to adapt or extend the traditional machine learning algorithms to handle multi-label datasets (MLD) directly [175]. The three most commonly used transformation methods are Binary Relevance (BR) [94], label powerset (LP) [176], and classifier chains (CC)[101]. BR transforms the multi-label problem into a set of independent binary problems. Then, each binary problem is processed by uisng a traditional classifier. Finally, the individual predictions are combined to get the subset of labels relevant to each test instance. LP considers each unique set of labels as class identifier, transforming the original MLD into a multi-class dataset. After using it to train a regular classifier, the predicted classes are back-transformed into the subsets of labels. Both BR and LP are the foundation for many multi-label ensemble-based methods. CC resolves the BR limitations by taking into account the label correlation task. The second approach, algorithm adaptation, focuses on introducing MLC-specific changes in classification algorithms. It includes approaches such as adjusting class weights or decision thresholds, modifying an existing algorithm, or constructing an entirely new algorithm to perform well on an imbalanced MLD. Several adaptations of the traditional classifiers have been proposed in the literature, such as Multi-Label k Nearest Neighbors (MLKNN) [177], multi-class multi-label perceptron (MMP) [178], and Ranking Support Vector Machine (Rank-SVM) [109]. A recent review of MLC methods is provided in [179].

The major challenge in a classification task is the data imbalance problem where it imposes difficulty in performing data analytics in almost all domains of research. In traditional classification, the imbalanced problem is well studied in recent times. An extensive review of methods for imbalanced data in single-label classification is already presented elsewhere [180]. The imbalanced problem in MLC is much more complicated than in single-label learning, as each instance can have multiple labels simultaneously. The imbalanced nature of an MLC includes the skewed distribution of the examples and their respective labels, that is, the labels are non-uniformly distributed over the data space. The problem transformation and adaptation approaches applied for the MLC task are not effective in handling the imbalance problem in an MLC. An imbalanced dataset, in general, becomes a significant challenge in many real-world applications, such as fraud detection [181], risk management [182], and medical diagnosis [183]. For example, in a disease diagnostic problem where the cases of the disease are usually rare as compared to the normal populations, the main interest of the task is to detect people with the diseases. Hence, an effective classification model is the one that could provide proper labeling of the rare patterns. The class frequencies in an imbalanced dataset can differ to a large extent, which affects the learning process of a classification model. The imbalanced class distribution has been extensively studied for traditional classification using the commonly existing approaches, such as resampling methods [180]. However, the existing methods cannot be directly applied as a solution to the imbalanced problem in an MLC due to an imbalance between labels and label-sets. The imbalance problem becomes even more complex for MLDs with a higher number of labels.

In this chapter, a literature survey was performed in order to identify a broad range of approaches for addressing the imbalanced problem in MLC. The contributions of this survey are threefold: (1) to the best of the authors' knowledge; it is the first survey study focused on the role of imbalance techniques in an MLC task. It presents the characteristics of an imbalanced MLD, a comprehensive survey of different approaches for imbalanced MLC and a summary of evaluation measures; (2) This chapter presents a comparative analysis of existing approaches and investigates the pros and cons of each approach; (3) The results presented here provide guidance for choosing the appropriate technique and developing better approaches for handling an imbalanced MLC in further studies in this area.

The rest of this chapter is organized as follows. Section 3.2 presents the

research methods and statistical trends. Section 3.3 describes the classification with imbalanced dataset, including its taxonomy and imbalanced level measures. Section 3.4 is the main section of this chapter, which discusses various approaches for addressing the imbalanced problem in MLC. Section 3.5 contains a short description of datasets and tools. Section 3.6 describes various metrics for the evaluation of the MLC model. Section 3.7 presents a comparative analysis of solutions with advantages and limitations. Finally, future research directions and conclusions are provided in sections 3.8 and 3.9, respectively.

## 3.2 Methods and Statistical Trends

In order to ensure as an objective selection of literature sources as possible, a well-defined search methodology for collecting source articles was adopted in this work. This methodology is presented in detail in this section.

### 3.2.1. Data Sources and Search Strategy

For this systematic review, research articles related to imbalanced MLC were searched for, in order to compile the published papers from 2006 up to 2019. First, well-known library databases that covered the different research fields were used as a source of information for searching and collecting the literature: DBLP, IEEExplore, Springer, ACM Digital Library, Elsevier, Science Direct, Google scholar, etc. Boolean operators were used for searching for terms with similar meanings and restricting the research. Predetermined search keywords that included a combination of query phrases, such as 'imbalanced multi-label classification' or 'addressing imbalanced problem' or 'multi-label dataset or 'multi-label prediction,' were included. We also attempted to search for articles from other sources (such as peer review journals and conferences).

### 3.2.2 Selection of Studies

The main focus of this paper is to review methods for handling the imbalanced problem in MLC. The following eligibility criteria, which had to be jointly satisfied, were used to select the relevant publications: (1) the study is based on imbalanced MLDs; (2) the work adopts or proposes methods for addressing imbalanced MLC; (3) experimental results evaluate MLC algorithms using multi-label measures; (4) the publication is a full-text article

written in English. Articles that provide an MLD-based evaluation of proposed approaches for addressing imbalanced MLC are selected for review without restrictions on the dates of publication. Published works with duplicated titles, abstracts, or content were manually removed, keeping only one copy of the publication. Generally, at the initial stage of searching, 392 publications were collected and identified, from which 86 were duplicates and 219 were discarded on the basis of title and abstract. Finally, by reviewing the full text of each paper, 74 papers were found to be relevant to this study.

### 3.2.3. Statistical Trends

Figure 3.1 presents the publication trends of imbalanced multi-label learning by plotting the number of publications from 2006 to 2019. The number of publications has shown stable growth for the years between 2012 to 2015 and 2016 to 2019 in comparison to the other periods. The number of publications was lowered in the year 2016 compared to 2015 and later showed an increase in the number of publications in the subsequent years. More recently, the number of published works on imbalanced MLC is much higher than the previous years. This suggests that the imbalanced MLC has remained a valuable research topic that has gained wide attention from researchers.

## 3.3 Classification With imbalanced Dataset

### 3.3.1 Imbalance in Single-Label Classification

In many real-world applications, it is common to have imbalanced numbers of learning samples for one class compared to the other class. In single-label classification with imbalanced data, at least one class is represented by only a small number of samples (i.e., the minority class) while the other class comprises the rest of the samples (i.e., the majority class). This problem is known as the class imbalance problem, which happens when the distribution of classes is not uniform among samples and results in a biased prediction of learning towards majority classes. The imbalance problem in single-label classification has been investigated in the literature; however, it is still considered as an open challenge in data analytics in many domains including medical diagnosis, customer-related data, churn prediction, text categorization, and fraud detection, where the class of interest is the minority class. Several kinds of literature in the machine learning community have shown

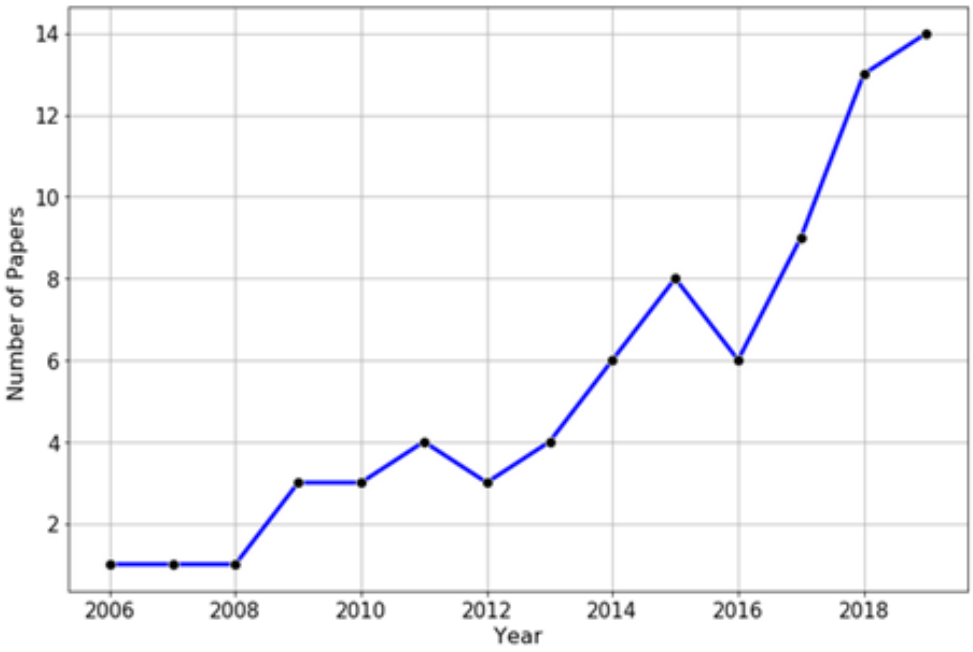


Figure 3.1: Publishing trends for Imbalanced Multi-label Classification

that learning with imbalanced data can leave the learned classifier with a performance bias; that is, a biased classifier exhibits very poor accuracy on the minority class but very high accuracy on the majority class [184, 185].

### 3.3.2 Imbalance in Multi-label Classification

In any classification task, the presence of imbalanced data [186] is a common and challenging problem that affects the learning process of a classification model. In particular, imbalance learning is a well-known and inherent characteristic of many MLDs, which affects the learning process of many classification algorithms. The imbalance problem in an MLD can be viewed from three perspectives: imbalance within labels, an imbalance between labels, an imbalance among the label-sets. In the case of imbalance within labels, each label usually contains an extremely high number of negative samples and a very small number of positive samples [187, 188]. In the imbalance between labels, the frequency of individual labels in the MLD is considered where the number of 1's (positive class) in one label may be higher than the number of 1's in the other label [190, 199]. Since every instance of an MLD is associated with several outputs or labels, it is common that some

of them are majority ones while others are minority labels, i.e., some labels have much more positive examples than others. The third type of label imbalance that usually occurs in MLD is the sparse frequency of label-sets [192]. If the full label-set is taken into account, the proportion of positive to negative examples for each class may be associated with the most common label-sets. In MLDs, due to the label sparseness, there are usually more frequent label-sets and unique label-sets. This also involves that some of the label-sets may be considered majority, and the remaining label-sets may be considered minority cases at the same time.

### 3.3.3. Characteristics of Imbalanced MLDs

This section discusses imbalance problems in multi-label datasets (MLDs) and characterization measures used for examining the characteristics of such datasets. In this and the following sections, we will use the following notation:
$M = \{(x_i, Y_i), i = 1, ..., m\}$ : an MLD consisting of $m = |M|$ multi-label examples,
$L = \{\lambda_j : j = 1, ..., q\}$: the set of all labels in the given multi-label classification problem,
$q$ : the number of labels, $q = |L|$
$x_i$ : the attribute vector of the $i^{th}$ sample in $M(i = 1, ..., m)$,
$Y_i \subseteq L$ : the actual label-set for the $i^{th}$ sample in $M(i = 1, ..., m)$,
$Z_i \subseteq L$: the predicted label-set for the $i^{th}$ sample in $M(i = 1, ..., m)$,
$ri(\lambda)$: the rank predicted by a label ranking (LR) method for the label $\lambda \in L$ . The most relevant label receives the highest rank, which is 1, and the least relevant one receives the lowest rank, which is $q$.

### 3.3.4. Characterization Measures in MLDs

Before building a classification model to solve a specific problem, we usually examine the characteristics of the dataset being studied to gain an understanding of the relationships between variables and to determine an appropriate model for it. When faced with an MLD, we must also examine the relationships between the labels, the concurrence level among imbalanced labels, and the imbalance level to determine the multi-labelness of the data. The most basic information which can be obtained from an MLD includes the number of samples, attributes, labels, and label-sets. The commonly used characterization measures of an MLD include label distribution measures, imbalance level measures, and concurrence level measures [193].

Figure 3.2 shows the taxonomy of characterization measures.

Label distribution measures: in some MLDs, the number of labels of each example is small, while in others, it is large. Each sample has an associated label-set, whose size can be in the range $\{0, ..., q\}$. There are two measures for evaluating the characteristics of an MLD, related to the distribution of labels: Cardinality (Card) and Density (Dens) [194]. Let $M$ be an MLD consisting of $m = |M|$ multi-label examples $(x_i, Y_i)$. Label cardinality of $M$ is the average number of labels of samples in $M$ (Eq.3.1). Label density is the average number of labels in $M$ divided by the number of all labels $q = |L|$ (Eq.3.2).

Imbalance level measures: most MLDs are imbalanced, in which some of the labels are very frequent, while others are quite rare. Therefore, it is important to define the level of imbalance in MLD, considering all the labels. Four different measures are proposed in the literature to assess label imbalance [195]: Imbalance ratio per label (IRLbl), Mean imbalance ratio (MeanIR), Maximum IRbl (MaxIR), and Coefficient of variation of IRLbl (CVIR).

$$Card(M) = \frac{1}{m}\sum_{i=1}^{m}|Y_i| \tag{3.1}$$

$$Dens(M) = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i|}{q} \tag{3.2}$$

Imbalance ratio per label (IRLbl) (Eq.3.3): let $M$ be an MLD with a set of labels $L$ and $Y_i$ be the label-set of the $i^{th}$ instance, IRLbl is calculated for the label $\lambda$ as the ratio between the majority label and the label $\lambda$. IRLbl is 1 for the most frequent label and a greater value for the rest. The larger the value of IRLbl, the higher the imbalance level for the concerned label.

$$IRLbl(\lambda) = \frac{\max_{\lambda' \in L}(\sum_{i=1}^{m} h(\lambda', Y_i))}{\sum_{i=1}^{m} h(\lambda, Y_i)}, h(\lambda, Y_i) = \begin{cases} 1, \lambda \in L \\ 0, \lambda \notin L \end{cases} \tag{3.3}$$

Mean imbalance ratio (MeanIR): it is the mean imbalance ratio among all labels in an MLD (Eq. 3.4).

$$MeanIR = \frac{1}{q}\sum_{\lambda \in L} IRLbl(\lambda) \tag{3.4}$$

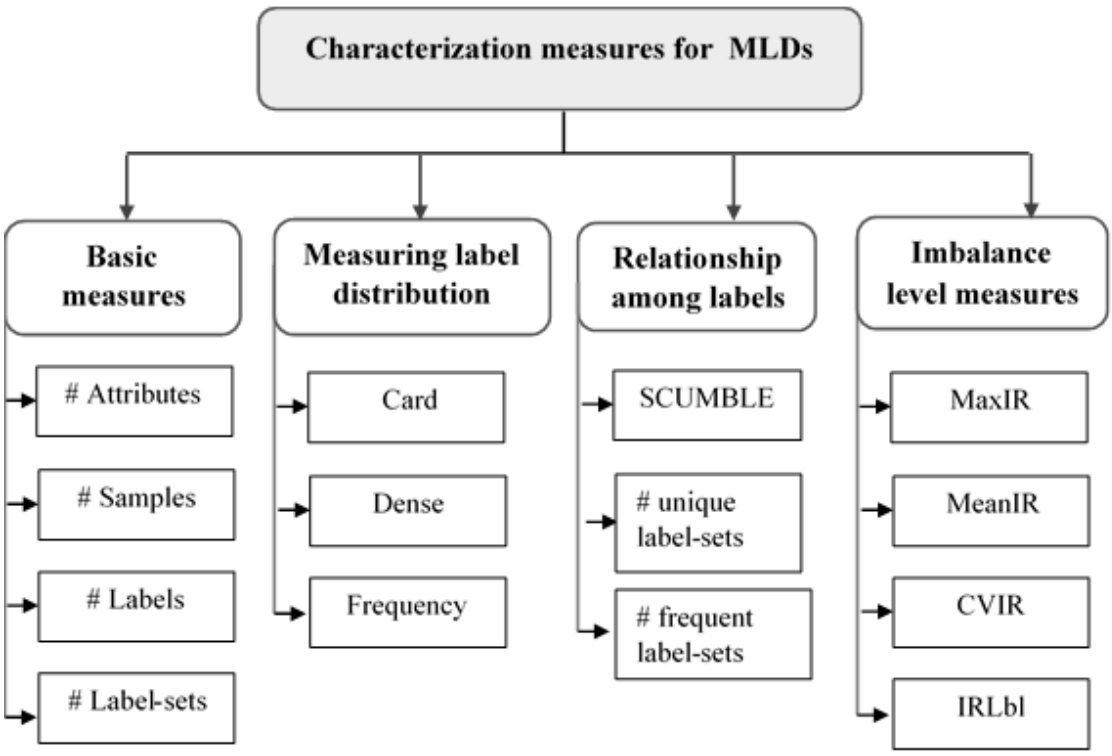


Figure 3.2: Taxonomy of characterization measures for MLDs

Maximum imbalance ratio (MaxIR): The ratio of the most common label against the rare one (Eq. 3.5).

$$MaxIR = \max_{\lambda \in L} IRLbl(\lambda) \tag{3.5}$$

Coefficient of variation of IRLbl (CVIR) (Eq. 3.6): CVIR measures the variation of IRLbl, i.e., the similarity of the level of imbalance between all labels. It indicates if labels experience a similar level of imbalance or, on the contrary, there are large differences among them. The higher the CVIR value, the higher would be this difference:

$$CVIR = \frac{IRLbl_{\delta}}{MeanIR}, \delta = \sqrt{\sum_{\lambda \in L} \frac{(IRLbl(\lambda) - MeanIR)^2}{q-1}} \tag{3.6}$$

Concurrence level measures: the number of different label-sets, as well as the amount of them being unique label-sets (appearing only once in MLD),

give us an indication of how sparsely the labels are distributed. The label-sets by themselves allow knowing how the labels in $L$ are related. SCUMBLE [202] is proposed to assess the concurrence among very frequent and rare labels. A small score will denote an MLD with not much concurrence among imbalanced labels, whereas a large one would evidence the opposite case. An MLD with a higher concurrence level would become more difficult to process using resampling algorithms.

## 3.4 Approaches for Imbalanced Multi-label Classification

The imbalanced approaches proposed for MLC can be divided into four categories: resampling methods, classifier adaptation, ensemble approaches, and cost-sensitive methods. Figure 3.3 summarizes the categorization of these approaches with descriptions in the next subsections. These approaches have been the most common strategies to rebalance the class distribution in the traditional classification (i.e., in single-label classification). They are adapted to the multi-label learning to reduce the imbalance problem between labels and among label-sets.

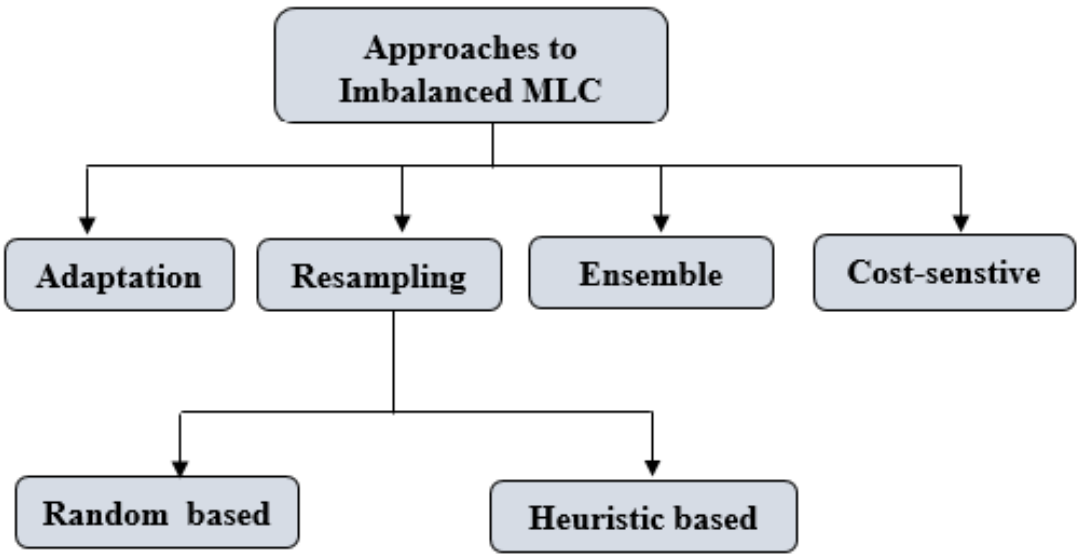


Figure 3.3: Categorization of methods proposed in the literature to address imbalanced MLC

### 3.4.1 Resampling Methods

The resampling approaches are the most commonly used techniques to handle imbalanced data. These approaches are based on the pre-processing of the MLDs. They aim to produce new, more balanced versions of MLDs, and they belong to the classifier-independent group. Resampling methods are based on undersampling [197], which removes samples associated with the majority label and oversampling [198], which generates new samples associated with the minority label, or both actions at the same time. The way in which the examples to be added or removed can also be grouped into two categories, random methods and heuristic methods. The former randomly choose the samples to be deleted or produced associated with a specific label. The latter can be based on disparate heuristics to search for the proper instances, as well as to generate new ones. The two resampling approaches have been adapted to deal with MLDs, as discussed in the next paragraphs.

**Multi-Label Random Resampling**: the random resampling method applied to MLC follows different approaches than the ones used in single-label classification, as the existing resampling methods cannot be directly used in MLC. These approaches can be based on the LP transformation, BR methods, imbalance measures, etc. LP-RUS and LP-ROS are two examples of resampling methods proposed in [195] based on LP transformation. The LP transformation method transforms the MLD into a multi-class dataset, processing each different combination of labels (label-set) as a class. LP-RUS randomly removes instances assigned with the most frequent label-set (i.e., a specific combination of labels), and the processing stops when the number of samples in the MLD is reduced by an indicated percentage. LP-ROS is a multi-label random oversampling method that works by cloning random samples of minority label-sets until the size of the MLD increases by the prespecified percentage. Although LP-based resampling has its own advantages to solve the imbalance problem, it is limited by the labels sparseness in the MLDs. In other words, there are MLDs with as many distinct label combinations as instances. This implies that all label-sets would be considered to be both majority and minority cases at the same time. Thus LP-ROS and LP-RUS could hardly fix the imbalance problem in such cases. An alternative approach to tackle this limitation would be evaluating the individual imbalance level of each label. ML-RUS and ML-ROS are examples of such approaches based on the frequency of individual labels, instead of the full

label-sets, isolating the instances with one or more minority labels [192]. The main aim of ML-RUS is to delete samples with majority labels and of ML-ROS to clone samples with minority labels. These two methods rely on IRLbl and MeanIR measures: Labels whose IRLbl is greater than MeanIR are considered to be minority labels, while labels whose IRLbl is smaller than MeanIR can be considered to be majority labels. ML-RUS determines what labels are the majority by means of their IRLbl value. One main limitation with these ML-based methods is that some of the minority samples selected by ML-ROS may contain the most frequent labels, due to the joint appearance of minority and majority labels. Therefore, the oversampling will include both the majority and minority labels. As a remedy to this problem, REMEDIAL is proposed in [199]. REMEDIAL method tackles the imbalanced problem by decoupling the majority and minority labels, of which the imbalance level is assessed by SCUMBLE. REMEDIAL could be either a standalone sampling method or can be combined with other resampling techniques, like REMEDIAL combines with MLSMOTE [199]. Other strategies, such as best first oversampling [200] and imbalance in hierarchical MLDs [201] have been utilized to address the imbalance problems of MLC.

**Multi-Label Heuristic Resampling**: in this approach, the instances to be deleted or cloned are heuristically selected, instead of being randomly chosen. The random resampling methods have problems in the loss of potentially useful information during undersampling, which brings overfitting in the course of oversampling. The heuristic approach can be employed as an alternative to overcome these limitations by selecting the right samples in the process of under sampling and oversampling. MLeNN is one of the heuristic multi-label undersampling approaches proposed by Charte et al. [202]. It is built upon the Edited Nearest-Neighbor (ENN) rule [203] and depends on the MeanIR and IRLbl measures to assess the imbalance level in MLDs. MLeNN is used to make a careful selection of instances to remove from the majority samples in a heuristic way and performs better than LP-RUS. MLTL is a similar heuristic-based approach recently proposed [204]. This method adopts the classic Tomek Link algorithm [205] to address the imbalance, which can be used as an under sampling or cleaning technique. Generally, heuristic-based under sampling methods, unlike random under sampling, try to eliminate the least significant instances of the majority class and thus minimize the risk of losing important information. However,

these methods also have some drawbacks: (1) they do not allow to influence the number of removed samples, which usually depends on the nature of the data; (2) they are difficult to apply when the minority and majority labels jointly appear in the same instances.

Heuristic-based synthetic instance generation has also been explored to handle imbalanced MLDs. A proposal in [206] is based on the original SMOTE algorithm [207] together with three transformation strategies. The first strategy uses a binary relevance method to transform instances into positive and negative to apply SMOTE. The second approach transforms instances in which minority label appears in isolation into positive and the remaining into negative. The third strategy considers all samples in which the minority label appears and applies SMOTE several times. In this paper [52], it was observed that the third method improved the results, whereas the other two produced a general degradation of performance. An extension of SMOTE, called MLSMOTE, applied to MLDs, was proposed in [208]. MLSMOTE considers a list of minority labels using the instances in which these labels appear as seeds to generate new instances. First, the nearest neighbors of the seed instances are found, and then the features of the synthetic instances are obtained by an interpolation technique. MLSMOTE takes into account several minority labels to produce synthetic instances instead of only one label, which is an advantage since most MLDs have multiple minority labels.

Another recently proposed approach is MLSOL [209]. This method focuses on analyzing imbalance by looking at the local characteristics of minority samples, rather than the imbalance of the whole dataset. MLSOL first calculates the weight vector for seed instance selection and a type matrix for synthetic instance generation based on the local label distribution. Once the seed instance is selected based on the weight vector, the reference instance is randomly chosen from the k nearest neighbors of the seed instance. An ensemble framework is incorporated into MLSOL to improve its robustness. The use of weighted sampling for seed instance selection and its ensemble version allows MLSOL to create more diverse models and to achieve better performance with greater error correction than MLSMOTE. In MLSMOTE, the labels of the synthetic instance are fixed, while in MLSOL, the labels of the new instance change according to its location, which avoids the introduction of noise. Other works are induction based under sampling [210] and reverse-nearest neighborhood-based oversampling [211].

In general, the resampling methods are popular approaches for dealing

with imbalanced data. However, since random oversampling usually involves exact copies of examples to increase the size of the data space, it may lead to overfitting [212, 213], and also requires more time during the training phase. Oversampling doesn't introduce new data, so it can not address the fundamental 'lack of data' issue. As a result, oversampling may not be effective at improving the detection of minority samples [213, 214].

### 3.4.2 Classifier Adaptation

Apart from resampling methods, adapting the existing machine learning algorithms is another way of facing the imbalance problem. Adaptation methods could be categorized as dedicated algorithms that directly learn the imbalance distribution from the classes/labels in the datasets. Some multi-label methods adapted to deal with imbalanced MLC have been proposed in the literature. In [215], a min-max modular network with SVM was proposed to address the imbalanced problem of MLDs. It works by decomposing a multi-label imbalanced classification problem into a series of small two-class subproblems. In the learning process, each subproblem can be trained by one of the standard classification algorithms, and then the outputs of the classifiers are combined by using minimization and maximization principles [216] to generate solutions to the original problem. This method works according to the principles of the Min-Max Modular network and presents different decomposition strategies to improve the performance of these networks.

Another proposal based on adaptation methods is presented in [217]. It uses an enrichment process in neural network training to address the multi-label and imbalanced data problems, such as semantic scene classification, robotic state recognition, and other real-world applications. The enrichment process manages the training data using three steps: the first step is an initialization, which uses a clustering method to group similar instances and gets a balanced representation to initialize the neural network. In the second step, the network is iteratively trained, as usual, while data samples are added and removed from the training set, according to their prevalence. The final phase checks if the enrichment process has reached the stop condition or it has to be repeated. This way, the overall balance of the neural network used as a classifier is improved.

Recently, an adaptation approach was proposed in [218] to address the imbalance in MLC. It is based on an asymmetric stage-wise loss function to adjust the loss cost of positive and negative samples dynamically. In [219],

imbalanced multi-modal multi-label learning (IMMML) was proposed. It was designed to tackle the imbalance problem in the subcellular localization prediction of the human proteins with multiple sites. The algorithm is based on a Gaussian process model, combined with latent functions on the feature space and covariance matrices to obtain correlations among labels. The imbalance problem is solved, giving each label a weighting coefficient linked to the likelihood of labels on each sample. Therefore, it is a very specific solution to a definite problem, hardly applicable in a different context.

The proposal in [220], Imbalanced multi-instance multi-label radial basis function neural networks (IMIMLRBF), is an extension of MIMLRBF [221]. IMIMLRBF is a multi-instance and multi-label classification algorithm based on radial basis neural networks. The adaptation works in two ways. First, the number of units in the hidden layer, with MIMLRBF being constant, is computed according to the number of samples of each label. Then, the weights associated with the links between the hidden and output layers are adjusted, biasing them depending on the label frequencies. In [222], an approach based on the multi-label hyper network was proposed to address the imbalance problem in MLC. In this algorithm, labels of an MLD are separated into two groups based on their imbalance ratios. These two groups are common labels and imbalanced labels. The algorithm works in two steps. In the first step, a multi-label hyper network is trained, and it produces preliminary predictions. In the second step, the correlations between imbalanced labels and common labels are used for refining the predictions obtained in the first step, thereby improving the classification performance.

Zhang et al. [223] proposed the class-imbalance aware algorithm named cross-coupling aggregation (COCOA). For each class in the dataset, COCOA combines the predictive results of a binary-class imbalance classifier corresponding to the current label and the predictive results of some multi-class imbalance learners. The final decision for each class label is obtained by aggregating the outputs of binary and multi-class learners. This approach has also been applied for a decision support system in medical diagnosis with imbalanced clinical data [224]. Pouyanfar et al. [225] propose recent work, entitled "multi-label multimodal deep learning framework for imbalanced data classification" to address challenges in multi-media data classification. The proposed framework handles the imbalanced problem in MLC by assigning a specific weight to each class automatically during the classification task. Other models based on a neural network in [226, 227]. Apart from

the above-mentioned methods, the RelieF feature selection algorithm [228], concept drift, and KNN based approach [229] have been employed to address the imbalance in MLC.

### 3.4.3 Ensemble Methods

Ensemble methods combine several base models in order to produce one optimal predictive model. The use of sets of classifiers as ensembles has proven to be effective in single-label classification. A similar approach has been used in MLC for improving predictive performance and solving the imbalanced problem. The ensemble of multi-label classifiers trains several multi-label classifiers. Thus, all the trained classifiers are different and can provide diverse multi-label predictions. There are several ways of joining the outputs of these classifiers [230]. An inverse random undersampling (BR-IRUS) method is proposed in [231]. BR-IRUS is implemented on an ensemble of binary classifiers that are trained for individual labels using a subset of the original data. The subset of the instances contains all samples in which the minority label is present, along with a small portion of the remaining samples. This way, each individual classifier solves a classification task. Joining the predictions given by the classifiers associated with a label, a more defined boundary around the minority label space is generated. In [232], a heterogeneous ensemble of multi-label learners is proposed by combining state-of-the-art multi-label methods. This method simultaneously tackles both the sample imbalance and label correlation problems. The ensemble is composed of five classifiers. All of them are trained using different algorithms on the same data. Several methods for joining the individual predictions are tested, along with different thresholding and weighting schemes with adjustments made through cross-validation.

The authors in [233] proposed an ensemble classifier called HPSLpred with an imbalanced source of human protein subcellular location prediction. HPSLpred integrates 12 kinds of basic classifiers to address the imbalanced problem. The authors in [234] used a two-stage stack-like ensemble of MLkNN classifiers to exploit label associations in MLC. The algorithm shows an improvement in comparison to MLkNN without stacking. ECCRU3 [235] extends the ECC resilient to class imbalance by coupling undersampling and improving the exploitation of majority samples. Furthermore, other ensemble classification algorithms have been employed in MLC, such as the ensemble of multi-label classifiers [236], bagging, and adaptive boosting [237].

### 3.4.4 Cost-Sensitive Approaches

Cost-sensitive methods use different cost metrics to describe the costs of any particular misclassified sample, aiming to minimize the total cost. Most commonly, these methods are applied to imbalanced learning by associating high misclassifying cost to the minority classes [238]. In traditional classification, the objective is to minimize the misclassification rate, and thus most classifiers assume that the misclassification costs are equal. A more general setting is the cost-sensitive classification, where the costs caused by different kinds of errors are not assumed to be equal. Cost-sensitive approaches can be incorporated both at the data level and at the algorithmic level, by considering higher costs for the misclassification of minority class samples with respect to majority class samples. In contrast to traditional classification, cost-sensitive learning studies in MLC are very few. The reason for this can be due to the fact that cost-sensitive learning strategies are difficult with regard to the assignment of an effective cost matrix. The cost matrix can be defined based on past experiences or domain experts with knowledge of the problem domain.

Some cost-sensitive approaches have been migrated to the multi-label scenario to explore a class-imbalance problem, among them: SOSHF [239] transforms the multi-label learning task to an imbalanced single label classification type via cost-sensitive clustering, and the oblique structured Hellinger decision trees address the new task. In [240], a cost-sensitive ranking support vector machine for MLD is attempted, which assigns a different misclassification cost for each label-set to effectively tackle the problem of imbalance in MLC. Another cost-sensitive multi-label learning is proposed in [241]. This work extends BR to consider the exploitation of the label correlations and exploration of the class-imbalance simultaneously. A cost-sensitive loss is utilized to tackle the class-imbalance problem.

## 3.5 Datasets and software tools

To evaluate the proposed methods of imbalanced MLC, most authors used publicly available benchmark MLDs with different formats (text, audio, images, etc.). The names of 26 MLDs in ARFF file format, along with their descriptions and statistics, are found from the online MULAN repository [242]. The MULAN repository is the most used resource by many authors of articles that concern the MLC task. Other MLD repositories include the

MEKA repository [243] and the R ultimate MLD repository [244]. The software tools associated with each repository are proposed in order to analyze MLDs and perform MLC. These include MULAN [245], MEKA [243], and multilearn library in Python [246].

## 3.6 Model Evaluation

Various metrics have been proposed in the literature to evaluate the classification performance of MLC models. Unlike the traditional classification, which produces a single class as output being either a correct or wrong prediction, the output of any multi-label classifier consists of a label-set predicted for each instance. The evaluation of MLC requires different measures with respect to the ground truth of multi-label prediction results. The measures can be broadly categorized into three groups: example-based [247], label-based [248], and ranking-based measures [194]. Example-based measures are computed individually for each example, then averaged to obtain the final value. Label-based measures are computed for each label, instead of per instance. The ranking-based metrics evaluate the ranking of labels with respect to the original MLDs.

### i. Example-Based Measures

**Hamming loss (HL)** is the most common performance measure in MLC, computed as the symmetric difference between the predicted and true labels and divided by the total number of labels in the MLD. The smaller the value of the Hamming Loss, the better the performance:

$$HL = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{q} \tag{3.7}$$

where $\Delta$ denotes the symmetric difference of the two sets and corresponds to the XOR operation in Boolean logic. HL measures the fraction of labels that are misclassified.

**Subset Accuracy (SA)** evaluates the percentage of correctly predicted labels among all predicted and true labels. This is a very strict evaluation measure as it requires the predicted set of labels to be an exact match of the true set of labels.

$$SA = \frac{1}{m}\sum_{i=1}^{m} I(Z_i = Y_i), \tag{3.8}$$

$$\text{where I(true)} = 1 \text{ and I( false)} = 0$$

**Accuracy** is the ratio of predicted correct labels with respect to the total number (predicted and actual) of labels for each instance.

$$Accuracy = \frac{1}{m}\sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{3.9}$$

**Precision** , computed as indicated in the equation 3.10, is the proportion of predicted correct outputs to the total number of predicted outputs, averaged over all instances.

$$Precision = \frac{1}{m}\sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{Z_i} \tag{3.10}$$

**Recall** measures the proportion of predicted correct labels to the total number of true labels, averaged over all instances.

$$Recall = \frac{1}{m}\sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{Y_i} \tag{3.11}$$

**F-measure** represents the harmonic mean of Recall and Precision, providing a balanced assessment between precision and recall. It is a weighted measure of how many relevant labels are predicted and how many of the predicted labels are relevant is obtained.

$$F - measure = 2 * (Precision * Recall)/(Precision + Recall) \tag{3.12}$$

As in single-label multi-class classification, the higher the value of accuracy, precision, recall, and F-measure, the better the performance of the learning algorithm.

**Hamming Score (Mean Accuracy)**: refers to the average of the accuracy for each label, where,$Z_j^{(i)}$ corresponds to the prediction given by the model and $Y_j^{(i)}$ to the real value . It is a label- wise accuracy, which rewards methods for predicting individual labels well [250].

$$Hamming\ Score = \frac{1}{m|L|}\sum_{i=1}^{m}\sum_{j=1}^{|L|}[Y_j^{(i)} = Z_j^{(i)}] \tag{3.13}$$

where [A] is an identity function, returning 1 if condition A is true

.

### ii. Label-Based Measures

Label based measures evaluate each label separately and then average over all labels. Therefore, any known measure, used for evaluation of a binary classifier (e.g., accuracy, precision, recall, F-measure, ROC, etc.), can be used here. Label-based measures are calculated for all labels by using two averaging operations, called macro averaging, where any of the measures can be computed on individual class labels first and then averaged over all classes and micro-averaging, the measures can be computed globally over all instances and all class labels. Let EM be one of the evaluation metrics, $FP_\lambda$ for False Positives, $TP_\lambda$ for True Positives, $FN_\lambda$ for False Negatives, and $TN_\lambda$ for True Negatives, the macro and micro averaged operations can be calculated as follows [251]:

$$EM_{macro} = \frac{1}{q} \sum_{\lambda=1}^{q} EM(TP_\lambda, FP_\lambda, TN_\lambda, FN_\lambda)$$

$$EM_{micro} = EM(\sum_{\lambda=1}^{q} TP_\lambda, \sum_{\lambda=1}^{q} FP_\lambda, \sum_{\lambda=1}^{q} TN_\lambda, \sum_{\lambda=1}^{q} FN_\lambda)$$

### iii. Ranking-Based Measures

**One Error** measures how many times the best-ranked label given by the model is not part of the true label-set of the example. The smaller the value of one error, the better the performance:

$$One\ error = \frac{1}{m} \sum_{i=1}^{m} \delta(\text{argmin } r_i(\lambda)), \delta(\lambda) = \begin{cases} 1, \text{if } \lambda \notin Y_i \\ 0, \text{otherwise} \end{cases} \tag{3.14}$$

**Coverage** is the metric that evaluates how far, on average, a learning algorithm needs to go down in the ordered list of predicted labels to cover all the true labels of an instance. Clearly, the smaller the value of coverage, the better the performance.

$$Coverage = \frac{1}{m} \sum_{i=1}^{m} \max(r_i(\lambda)) - 1, \lambda \in Y_i \tag{3.15}$$

**Ranking loss** (RL) measures how many times a relevant label appears ranked lower than a non-relevant label. The smaller the value of RL the better the performance:

$$RL = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i||\overline{Y_i}|}|E| \tag{3.16}$$

where $\overline{Y_i}$ is the complementary set of $Y_i$ with respect to $L$ ,
$E = \{(\lambda_a, \lambda_b) : r_i(\lambda_a) > r_i(\lambda_b), (\lambda_a, \lambda_b) \in Y_i \times \overline{Y_i}\}$

**Average precision (AvgPrec)** evaluates the average fraction of labels ranked above a particular label $\lambda \in Y_i$ which actually are in $Y_i$. The higher the average precision, the better the performance.

$$AvgPrec = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{\lambda \in Y_i}\frac{|\{\lambda' \in Y_i : r_i(\lambda') \leq r_i(\lambda)\}|}{r_i(\lambda)} \tag{3.17}$$

## 3.7 Comparative Analysis

In this section, we present the comparative analysis of the different methods proposed in the literature for addressing an imbalance problem in MLC. Table 3.1 depicts the advantages and disadvantages of classifier adaptation, resampling, ensemble, and cost-sensitive methods. These methods are effective alternatives for imbalanced MLC tasks. However, there exist various constraints with respect to addressing the imbalance. The adaptation method makes the model insensitive to the imbalanced sample distribution by modifying the base classifier. This method is inefficient when the label space is too large and requires extensive knowledge about the base classifier and the problem domain. Resampling methods proposed for imbalanced MLC are advantageous as they are classifiers independent and that do not require any specific multi-label classifier to preprocess MLDs. Thus, a preprocessed MLD can be used as input to any of the MLC algorithms. However, the large differences in imbalance levels between labels and the high level of concurrence among imbalanced labels would greatly influence the behavior of resampling approaches, and, as a result, only certain MLDs with the lowest concurrence level can be most benefitted from the resampling. Ensemble approaches have a problem of computational complexity since more than one classifiers have to be trained and combined to obtain a final prediction result.

Table 3.1: Advantages and disadvantage of different categories

| **Approaches** | **Advantages** | **Constraints/Disadvantages** |
|---|---|---|
| Resampling Methods | Can be applicable to any MLC classifier.<br>Classifier independent | High level of concurrence between imbalanced labels and a large number of unique label-sets.<br>May introduce noisy data. |
| Classifier adaptation | Effective in a certain context.<br>Original data will not be affected | Requires extensive knowledge of the specific classifier and problem domains. |
| Ensemble methods | Decrease variance and improve prediction.<br>Reduce overfitting | Computational complexity.<br>No clear criteria for selecting the type and number of MLC classifiers. |
| Cost-sensitive | Computationally efficient. | Real cost values are unknown in most applications domains |

Table 3.2: Comparison of specific methods proposed for addressing imbalanced MLC

| **Article** | **Approaches used** | **Balancing Method** | **Advantages** | **Limitations** |
|---|---|---|---|---|
| (Charte et al. 2013) [195] | LP-based [16] | Random resampling | Helps to reduce imbalance among the label-sets | label sparseness in the MLD, hardly fix the imbalance problem |
| (Giraldo-Forero et al. 2013) [206] | BR based [89] | Heuristic based | Easy to apply SMOTE for a class-imbalance problem | It doesn't consider imbalance between labels/label-sets |
| (K. Chen et al.,2006) [215] | Decomposition strategy [97] | Classifier adaptation | Subproblems can be balanced | One label may happen more frequently than other |
| (Charte et al. 2015) [192] | imbalance measure of individual labels | Random resampling | Reduces highly imbalanced labels | The joint appearance of the majority and minority labels affect one another |
| (M. Tahir et al.,2012) [231] | BR based | Ensemble approach | Reduces class-imbalance problem | Doesn't consider imbalance between labels/label-sets |
| (Pereira et al,2019) [204] | LP based | Heuristic based | defined threshold for Hamming distance to remove majority label | Difficult to apply to highly concurrent imbalanced labels |
| (F. Luo et al. ,2019) [218] | asymmetric stage-wise functions | Classifier adaptation | Accuracy on minority samples can be improved | More applicable to only missing (unlabelled) labels [98] |
| (M. A. Tahir et al.,2012) [232] | LP, BR and CLR based transformation | Ensemble approach | Tackles both class imbalance and label correlation | Computationally intensive and base classifiers may be problem-specific |
| (Charte et al.,2019) [199] | BR and LP | Heuristic and random resampling | Solves the limitation of multi-label oversampling | Limited to high label concurrence problem under certain conditions |
| (Ding et al,2018)[255] | BR and CC | Cost-sensitive | Use of penalty function to balancing | Does not consider imbalance among labels |

In Table 3.2, a detailed comparison of various methods from the different approaches using various parameters is presented. The comparison criteria include the MLC approaches used, balancing method, advantages, and limitations of each proposed approach. The authors in [204] used MLC algorithms to compare the state of the art multi-label resampling approaches using different imbalanced MLDs with varying level of imbalance. Table 3.3 presents the experimental result of resampling approaches on six MLDs using a micro F-score as an evaluation metric and RAkEL as an MLC classifier. The experimental results in Table 3.3 indicate that LPROS achieved better results on three datasets (emotions, scene, and yeast), while MLROS and MLTL have shown better results on CAL500 and Medical datasets, respectively.

Table 3.3: Comparison of resampling methods for MLC with imbalanced datasets

| **Approaches** | **Datasets** | | | | | |
|---|---|---|---|---|---|---|
| | CAL500 | Emotions | Enron | Medical | Scene | Yeast |
| None | 0.3354 | 0.621 | 0.5496 | 0.8132 | 0.6237 | 0.5812 |
| LPROS [195] | 0.4924 | **0.6814** | 0.6306 | 0.8761 | **0.762** | **0.6721** |
| LPRUS [195] | 0.3751 | 0.5838 | 0.5158 | 0.7853 | 0.6339 | 0.5823 |
| MLROS [192] | **0.5413** | 0.6395 | 0.6694 | 0.8354 | 0.65 | 0.6671 |
| MLRUS [192] | 0.3255 | 0.5846 | 0.5259 | 0.8345 | 0.6919 | 0.5677 |
| REMEDIAL [256] | 0.2951 | 0.325 | 0.1135 | 0.637 | 0.5648 | 0.456 |
| REMEDIAL-HwR-ROS [199] | 0.2503 | 0.5111 | 0.2822 | 0.5064 | 0.6361 | 0.3929 |
| REMEDIAL-HwR-HUS [199] | 0.1293 | 0.349 | **0.684** | 0.7841 | 0.7176 | 0.3849 |
| REMEDIAL-HwR-SMT [199] | 0.1542 | 0.3056 | 0.1851 | 0.4114 | 0.3888 | 0.3772 |
| MLeNN [202] | 0.3466 | 0.621 | 0.6489 | 0.8774 | 0.6415 | 0.5846 |
| MLSMOTE [208] | 0.3839 | 0.4265 | 0.6125 | 0.8546 | 0.4532 | 0.5794 |
| MLTL [204] | 0.372 | 0.6409 | 0.6499 | **0.8798** | 0.7502 | 0.6348 |

## 3.8 Future Work

Existing works have proposed various techniques to tackle the imbalanced issues of MLC. However, several challenges remain, and imbalanced classification from the MLD still requires significant development. The following are some of the possible future research directions to deal with imbalanced MLC.

1. The success of currently available multi-label resampling algorithms is highly influenced by many factors:
   (i) Joint occurrence of minority and majority labels in the same instance. The potential existence of samples associated with rare and frequent labels in an MLD could make the resampling strategies ineffective. A recent study in [199] has attempted resampling by decoupling imbalanced labels, but it has limitations for some resampling methods (e.g., MLeNN) and MLC algorithms (e.g., MLkNN) due to that the decoupled instances are located in the same position. More sophisticated approaches are needed, considering label-set based relocation and defining thresholds for decoupling that can be able to work with any of the available MLC methods.
   (ii) MLDs with a large number of imbalanced labels pose a scalability challenge. Some methods have been proposed by Wang [229] to solve this problem. This approach could reduce Hamming Loss but did not completely eliminate it. Some approaches, such as parametrization and embedding [252], may help to address such challenges.

2. In MLC, there is a need for imbalance–aware classifiers that do not require resampling strategies. It seems promising to use the existing MLC methods (such as hierarchical MLC or classifier chains) and combine them with the imbalance–aware solutions that are available in the multi-class classification domain. An ideal goal would be the development of such multi-label classifiers that display similar performance to canonical methods on balanced multi-label problems while being at the same time robust to the presence of imbalance.

3. Ensemble methods are well-known to tackle both imbalance and label correlation problems. They are advocated by employing the non-trainable average combining rule. However, since MLC algorithms are computationally intensive and MLDs are highly imbalanced, it opens new research challenges on how to use other combination techniques

efficiently, such as trainable combiners (fuzzy integral) [253] or class indifferent combiners (decision templates and Dempster-Shafer combination) [254]. The other issue that needs further investigation is how to select the base classifiers in MLC since different combinations of base classifiers may perform differently for the specific problem domain. Moreover, there are no clear indicators of how large the ensemble should be constructed when applied to MLC tasks.

4. The other strategy can be the use of hybrid methods, concentrating on combining previously mentioned approaches to take advantage of their strong points and reduce their weakness. It is recommended to combine one of the resampling methods with another one or with adaptation methods investigating the potential to improve the results in these cases.

5. Another interesting direction is to investigate the possibilities of using cost-sensitive learning solutions. RAkEL [173] is the most popular method, transforming a multi-label problem with a large number of label-sets into smaller subsets. Hence, it seems straightforward to balance label-set distribution by automatically generating a misclassification cost vector in accordance with the label-set distribution.

6. Many of the proposed approaches to address imbalanced MLC in literature are based on first-order frameworks, such as binary relevance (BR) method. However, further research is required to extend these approaches to the second and higher-order frameworks to take into account label correlations. This can be approached by using information about label correlations along with other strategies to improve the learning performance of extremely imbalanced label distributions.

## 3.9 Conclusions

This chapter presents the first survey of handling the imbalanced problem in multi-label data classification (MLC), which includes the characteristics of the data, problem descriptions, solutions, and limitations of the approaches for solving an imbalanced problem. In this study, numerous articles related to imbalanced MLC published between 2006 and 2019 were collected and reviewed. Various methods and techniques that have been proposed to overcome the negative impacts of imbalanced MLC can be grouped into four categories: resampling methods, classifier adaptations, ensemble methods,

and cost-sensitive learning approaches. These approaches have their own limitations, even though they have shown achievement in handling imbalanced classes, labels, and label-sets in MLC. For example, methods which are proposed for handling imbalance problem between labels cannot be applied to handle the imbalance problem among label-sets. We also found that research in imbalanced MLC is very limited and that the majority of the existing works addressing the imbalance problem focus on single-label classification. Despite a growing demand for multi-label classification in different domains, developing a comprehensive framework for handling an imbalanced problem in an MLD is still understudied. As a result, this chapter concludes with a discussion on the challenges of imbalanced MLC and some future research directions that are worthy of further study.

# Chapter 4

# Predictive Modeling for Frailty Conditions in Older Adults

In this chapter, frailty risk predictive models were developed based on the whole elderly population of the Piedmont region, Italy.The predictive models were designed to detect and predict frailty according to the expected risk of various adverse health outcomes (mortality, urgent hospitalization, disability, fracture, and emergency admission at the emergency department) within 12 months.The models were developed based on administrative health data containing about 1 million elderly people aged 65 or older with 58 input variables and 6 output variables. First, six problems/outputs were identified as surrogates of frailty. Then, the imbalanced nature of the data was resolved through resampling process and a comparative study between the different machine learning algorithms – Artificial neural network(ANN), Support vector machines (SVM), Random Forest (RF), Logistic regression (LR), Decision tree (DT), and Genetic programming (GP) – was carried out. The performance of each model was evaluated using a separate test dataset through both the holdout and ten-fold cross-validation methods.

## 4.1 Introduction

The elderly population has been conventionally defined as a chronological age of 65 years or older [257], and frailty is frequently mentioned in studies related to the elderly population [7].The health condition of frail people is characterised by several diagnoses, complex medication routines, frequent physician visits and hospitalisations, impaired activities of daily living, cognitive decline and reliance on a caregiver for support. Frail older people

may also suffer from isolation and depression, which compounds their health problems and may influence non-adherence to prescribed medical regimes [258]. Recognition of frailty or pre-frailty is important for clinical practitioners and also policy-makers, as it poses a greater risk of adverse health outcomes such as falls, increased morbidity, physical and psychosocial dependence and death [259].

Frailty in elderly people was first characterized as a physical phenotype by Fried et al.[20]. According to this study, frailty is defined on the basis of five physical components: exhaustion, weight loss, slow gait speed, weakness, and low levels of physical activity. People who meet three or more of the above mentioned physical components are classified as frail. Those people who meet one or two criteria as pre-frail and people who meet none of these criteria are classified as not frail. This research was only phenotypic and didn't consider other causes, such as psychological and cognitive factors, to measure frailty. On the other hand, Rockwood et al. [21] developed a model to detect frailty. It is based on a comprehensive geriatric assessment and takes into account different parameters of symptoms, signs, abnormal laboratory values, disease states, and disabilities, referred to as deficits, to define frailty. In [260], the comparison of the frailty phenotype and the frailty index models were also widely discussed. A retrospective study based on a logistic regression model was proposed in [261] to develop frailty risk index and validate their content using health record data. There are also few models that are derived from a single source of information, like primary care electronic health record data and only insurance claims data [262]. More recent work on frailty was proposed by F. Bertini et al.[263] using logistic regression. In this paper, they proposed a frailty prediction model using a broad set of socio-clinical and socioeconomic variables. Their model was designed to detect and categorize frailty according to the expected risk of hospitalization or death. In general, the frailty indexes proposed in most literature have focused on the possible risk factors associated with frailty in the elderly population, but predicting who is at risk of frailty problems is still requires further investigation.

Several scales and models have been proposed for the detection of frailty [264]; however, a precise operational definition of frailty or a standard method for its screening and diagnosis is still lacking [265]. Moreover, each of the available tools intended to detect frailty poorly agrees with each other when applied to the same population [266]. In different clinical settings where the standard measure of frailty is missing, and the care of the

elderly is a priority, it is imperative to have a specific predictive model in the prediction of frailty according to the characteristics of the population being studied. Therefore, the aim is to detect multiple outcomes of frailty (mortality, disability, fracture, hospitalizations, and emergency admissions) using large administrative health databases of elderly people in Piedmont, Italy.

This chapter focuses on exploring the existing machine learning techniques (artificial neural network, genetic programming, support vector machines, logistic regression, random forest, and decision tree) to predict frailty according to the different adverse health outcomes. These approaches were considered for their performance and practical usefulness in the analysis of different types of medical data.

## 4.2 Methods

### 4.2.1 Data Source

This study is based on the Piedmontese Longitudinal Study. The data is collected using an individual record linkage that is available for about four millions of Piedmont (Italy) inhabitants between the Italian 2011 census and the administrative and health databases (enrollees registry, hospital discharges, drug prescriptions, outpatient clinical investigation database, and health exemptions) and that is included in the Italian Statistical National Plan. Subjects aged 65 years and above are included in the study. The dataset contains 1,095,612 subjects and 64 different variables (58 input and 6 output variables). The dataset includes a wide variety of predictor variables, including clinical and socioeconomic aspects. All outcomes and comorbidity variables are represented by Boolean values. The demographic variables such as age, marital status, citizenship, education level, income status, family size, and others are specified using the dummy variables. The 'age' variable is grouped into six categories, with 65-69 used as the first category. The output variables are described as outcomes or measurable changes in the health status of patients. The output variables include mortality, disability, urgent hospitalization, fracture, preventable hospitalization, and accessing the emergency department (ED) with red code. The color codes assigned to patients may vary from one hospital to another, but in this study, a red code is used to identify patients with severe symptoms that need immediate care. Since we intend to develop predictive models for

these frailty indicators, we extracted as input data those collected in 2016, while using as output values those collected in 2017.

Table 4.1: Description of output variables in the dataset.

| **Variables** | **Category** | **Code** | **Number** | **Percent (%)** |
|---|---|---|---|---|
| Mortality | No | 0 | 1,053,790 | 96.18 |
| | yes | 1 | 41,823 | 3.82 |
| Access to ED[a] with red code | No | 0 | 1,088,124 | 99.32 |
| | yes | 1 | 7,489 | 0.68 |
| Disability | No | 0 | 1,064,186 | 97.13 |
| | yes | 1 | 31,427 | 2.87 |
| Fracture | No | 0 | 1,088,530 | 99.35 |
| | yes | 1 | 7,083 | 0.65 |
| Urgent hospitalization | No | 0 | 1,056,695 | 96.45 |
| | yes | 1 | 38,918 | 3.55 |
| Preventable hospitalization | No | 0 | 1,076,541 | 98.26 |
| | yes | 1 | 19,072 | 1.74 |

ED[a]: emergency department

The way the data set is organized is such that one patient can have multiple outcomes. Such type of data is what we call 'multi-output' or multi-label dataset. For simple implementation and analysis, this multi-output dataset is transformed into six single-output problems associated with each output variable. Decomposing the original data into six independent datasets helps to study each output independently for the given number of similar risk factors. Transforming the original problem into single independent problems is a straightforward way to implement using classical machine learning methods. Additionally, with this method, we can take full advantage of such algorithms which consider learning problems that contain only one output.

The six different problems that are associated with frailty conditions are considered separately in the analysis, which results in six independent binary classification problems. All the input variables used in the study are presented in Appendix A. Table 4.1 contains descriptive statistics for all output variables, where the frequency distributions of each category of an output variable are represented as counts and percentages. Table 4.1 clearly shows how the dataset is, for each output variable, unbalanced.

In fact, there are approximately 4% records that have mortality risk as '1', and the rest 96% have mortality risk as '0'. Similarly, there are only approximately 1%, 1%, 2%, 3%, and 4% of the records, which have risk as '1' for access to the emergency admission with red code, fracture, preventable, hospitalization, disability, and urgent hospitalization, respectively. This is clearly an indication of an imbalanced dataset, as the number of subjects from the positive sample is much smaller than the number of subjects of the negative sample.

Most machine learning techniques may suffer from such extremely unbalanced datasets, and, as a result, they may be biased towards the majority class. Instructing a model with an algorithm that tries to maximize the accuracy will naturally lead to classifying everything as the major class and does not give acceptable results. Therefore, it is important to balance the data before developing the machine-learning model.

### 4.2.2 Handling Imbalanced Dataset

Imbalanced data sets are common in medicine and other domains. The issue of imbalanced datasets has gathered wide attention from researchers during the last several years [180]. It occurs when the samples represented in a problem show a skewed distribution, i.e., when there is a majority (or largest number of negative samples) and a minority (or least number of positive samples) in a dataset. Analyzing such a complex nature of the dataset becomes an issue in the machine learning community, including genetic programming, and it is observed that most of the traditional machine learning algorithms are very sensitive with imbalanced data. Usually, accurate classification of minority class samples is more important than majority class samples, especially in medical diagnosis. Table 4.1 presents an imbalanced data in each problem (mortality, access to ED with red code, disability, fracture, urgent hospitalization, and preventable hospitalization). The imbalanced proportions between the positive and negative classes of the six datasets are treated independently. Providing imbalanced data to a classifier will produce undesirable results such as much lower performance and increasing the number of false negatives.

There are various approaches to deal with imbalanced data that have been used in the literature, such as resampling and cost-sensitive learning methods [267]. In this study, we choose the resampling methods, which are based on under-sampling [197] and oversampling [268].These methods are

advantageous because they are classifier independent and can be used as a pre-processing step, in which the processed data can be given as input to any classifier. Oversampling is the process of replicating samples from the minority class to balance the data. The limitation of oversampling is that it may cause an overfitting problem as it clones the same instance and requires more time to execute as compared to the under-sampling approach. As a result, it is recommended to use it when the dataset is quite small in size. Another issue with oversampling is that as our aim is to detect minority classes, oversampling changes the class that we want to identify, which may not be acceptable in some critical real-time problems [269]. Under-sampling balances the imbalanced data by reducing the size of samples from the majority class. One limitation of the under-sampling approach is that it may lead to loss of important information or bias in the data. From a practical point of view, some literature showed that under-sampling tends to outperform oversampling in some settings [270], while others demonstrate that oversampling performs better than under-sampling. In high-dimensional data, oversampling performs poorly than oversampling [271], while under-sampling performs poorly in very small datasets. In our case, since the amount of collected data is sufficient, we adopted under-sampling to rebalance the sample distribution followed by a statistical test to avoid bias and ensure representativeness between samples. Since our data is multi-output data, we followed these simple steps to obtain balanced and independent datasets:

1. Filter all positive and negative samples from the original dataset based on the values of the output variables. Samples with at least one positive class value from the six outcomes are grouped as a positive sample, which accounts for 10% of the original dataset, and all remaining are grouped as a negative sample, which comprises 90% of the original dataset.

2. Keeping all the 10% samples in the positive class (minority group), randomly select an equal number of samples (10%) in the negative class (majority group).

3. Check whether the randomly selected 10% negative samples are representative of the remaining negative samples (90%). After checking that the test is reasonably significant, we obtained a new multi-output dataset of size 211924 each. A statistical test was applied in all variables to decide whether the distribution of frequencies of a variable in

the 10% sample was representative of the same variable in the 90% sample. Since all the variables in the study are categorical, we used a chi-square independence test with a significance level of 0.05 to check if there is a significant difference between the 10% sample the 90% sample with respect to input variables. The yielded chi-square statistic and the P-values are assessed to support the significance of the test's conclusion. The results of the chi-square test between 10% and 90% negative samples are shown in Appendix B.

4. Once the test is significant, we decomposed the multi-output dataset into six independent datasets. Then, an equal number of positive and negative samples are selected randomly from each dataset.

### 4.2.3 Predictive Models

The machine learning (ML) approaches selected for this study are support vector machines (SVM), artificial neural networks (ANN), random forests (RF), decision trees (DT), logistic regression (LR) and genetic programming (GP). A brief summary of these learning algorithms is presented below:

SVM is a robust classifier that was originally designed to identify two classes that require a huge amount of training data to select an effective decision boundary (Figure 4.1). There exist several works on prediction and classification using SVM [272].

SVM converts the original feature space into a higher dimensional feature space based on a kernel function and then obtains support vectors to maximize the separation (margin) between two classes. SVM first approximates a hyperplane for segregating the two class labels. Accordingly, SVM takes samples from both the classes, named as support vectors, which are closest to the hyperplane. The total distance between the hyperplane and its support vectors is called the margin. SVM then iteratively aims to optimize and/or maximize the margin between the hyperplane and supports vectors, thereby finding the most generalizable decision boundaries. When the dataset cannot be linearly separable, certain kernels are implemented in the SVM to appropriately transform the feature space into higher dimensional and convert back into the original feature space. This is called kernel function. There are various kernels and parameters that are used to improve the performance of classification by SVM [273]. In this study, the radial basis function (RBF) kernel is used with different values of gamma and the

regularization parameters for solving the binary classification problem. It has been used to predict negative health outcomes or events in frailty data by plotting the training dataset where a hyperplane classifies the points into two classes, presence and absence of frailty.

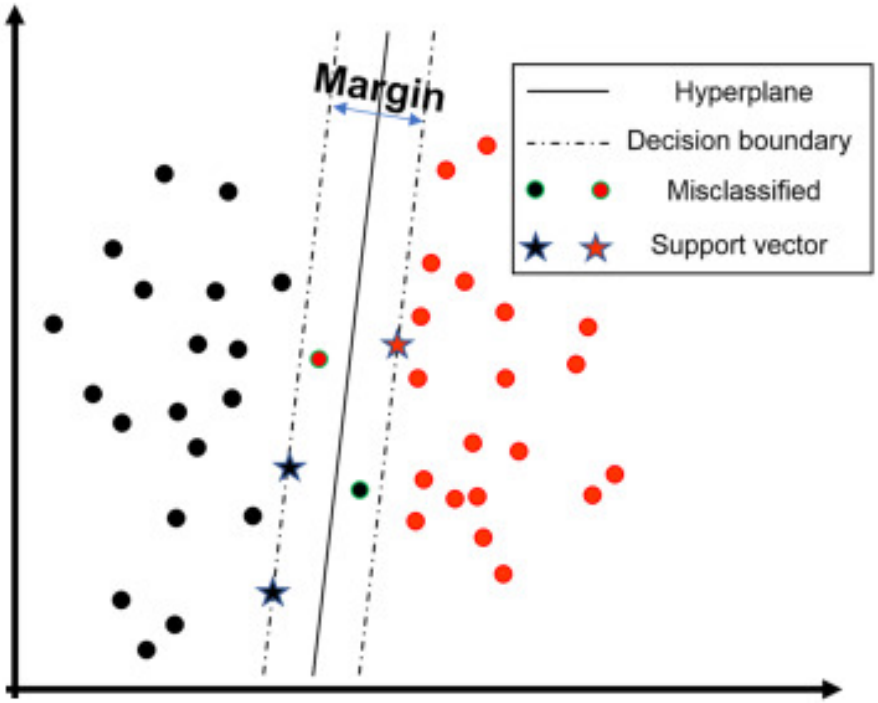


Figure 4.1: A typical SVM classifier without kernel function on a dataset that has two features and two classes. All training samples are represented as circles or stars. Support vectors (denoted as stars) are from the training samples such that they are closest to the hyperplane among the other training samples for each of the two classes. Two training samples have been misclassified because they lie on the wrong side of the hyperplane.

Another analytical technique used in this study is artificial neural networks (ANNs), which have been successful in solving classification problems in different domains [274]. Based on the functioning of biological neural networks, ANNs are dense networks of interconnected artificial neurons that get activated based on inputs. Among the various types of ANNs, in this thesis, we used multilayer perceptrons neural network (MLPs) with back-propagation learning algorithms. MLP, the ANNs most commonly used for a wide variety of problems, is based on a supervised procedure, i.e., the network develops a model based on samples in data with known outputs.

An MLP contains three layers (input, hidden, and output) with nonlinear computational elements (also called neurons and processing units). A neuron, often called a node or unit, is the basic unit of computation in a neural network. The information flows from the input layer to the output layer through the hidden layer (Figure 4.2). All neurons from one layer are fully connected to neurons in the adjacent layers. These connections are represented as weights (connection intensity) in the computational process. The weights play an important role in the propagation of the signal in the network. They contain the knowledge of the neural network about the problem-solution relation. The number of neurons in the input layer depends on the number of input features in the model, whereas the number of neurons in the output layer is equal to the number of output variables. The number of output neurons can be single or multiple. In frailty modeling, input features are generally given to the input layer as independent variables to predict outcome variables associated with frailty, which are given to the output layer as target values corresponding to the given input values. Moreover, both the numbers of hidden layers and their neurons are dependent on the complexity of the model and are important parameters in the development of the MLP model. The main parameters used in MLP, which include activation function, solver, hidden layer size, and learning rate, are configured for the classification work.

An MLP is trained to minimize errors between the desired target values and the values computed from the model. If the network gives the incorrect answer, or if the errors are greater than a given threshold, the weights are updated to minimize them. Thus, errors are reduced, and, as a result, future responses of the network are likely to be correct. In the learning process, datasets of input and desired target pattern pairs are presented sequentially to the network. The learning algorithm of an MLP involves a forward propagation step followed by a backward propagation step. MLP algorithms are widely discussed in the literature [275].

We also explored the potential of tree-based classifiers (Decision trees and Random forests) for the prediction of outcomes in each frailty problem. Decision trees (DT) builds classification models in the form of a tree structure. DT algorithms are effective in that they provide human-readable rules of classification. The main algorithms used in decision trees are ID3, C4.5, and CART [276], which build decision trees using the concept of information entropy. In our study, the CART algorithm is used for building the decision tree with hyperparameters set for each problem. Random forests

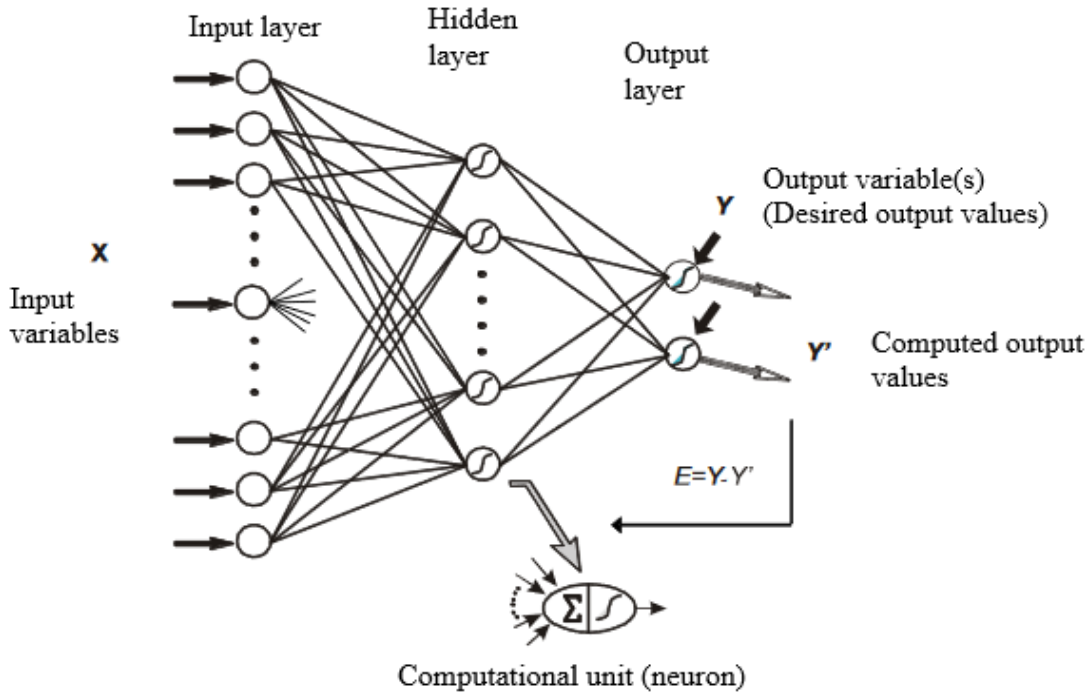


Figure 4.2: A typical artificial neural network architecture with three layers. $X$, input variables; $Y$, output variables; and $Y'$, values computed from the model.

(RF) consist of a large number of individual decision trees that operate as an ensemble, and a bootstrap sample technique is used to train each tree from the set of training data. Each tree gives a classification, and the forest chooses the classification having the most votes (over all the trees in the forest). For the final decision, RF classifier aggregates the decisions of individual trees; consequently, RF classifier exhibits good generalization.It seems that using RF results in increased stability as compared to using single decision trees. Tree-based classifiers are known for the prediction task in the medical domain [277]. The set of hyper-parameters, such as the number of trees in the forest, the maximum number of features considered for splitting a node, the maximum number of levels in each decision tree, etc. have been set for each problem.

Logistic Regression (LR), a specific type of multivariate regression, is the most common and well-established binary classifier [278]. LR is used to model only a dichotomous variable, which usually represents the presence or absence of an outcome or event based on a set of predictor variables. It

predicts an event of occurrence by fitting a dataset into a logit function. LR analysis has also been used particularly to investigate the relationship between binary or ordinal response probability and explanatory variables. In this study, like other ML models, LR has been used to distinguish frail and non-frail subjects.

Genetic programming (GP) has also been applied to the prediction of adverse outcomes asscocited with frailty. GP is typically designed to address the problem of automatic program synthesis and automatic programming. GP accomplishes this task by generating a population of computer programs over many generations using operations of natural selection [279]. GP is a search and optimization algorithm that iteratively transforms a population of computer programs into a new generation of programs using various genetic operators. The most commonly used operators are crossover, mutation, and reproduction. The crossover operator recombines randomly chosen subtrees among the parents and creates a new program for the new population. The mutation operator replaces randomly chosen subtree by a randomly generated tree, while the reproduction operator replicates a selected individual to a new population.The general problem solving process of GP includes the following five steps (Figure 4.3).

1. Generate an initial population of functions and terminals of the problem (computer programs) randomly. Each of the randomly generated computer programs is considered as candidate solutions to the problem.
2. Execute each program in the population and give it a fitness value according to how well it solves the problem.
3. Create a new population of computer programs, i.e, the next generation is produced using genetic operations:
   (a) Copy the best existing programs based on fitness value
   (b) Create new computer programs by mutation.
   (c) Create new computer programs by crossover
4. Steps 2 and 3 will be repeated until a termination criterion is matched, which can be finding the best program or reaching the maximum number of generations.
5. The best computer program that is shown in any generation is designated as the result of GP.

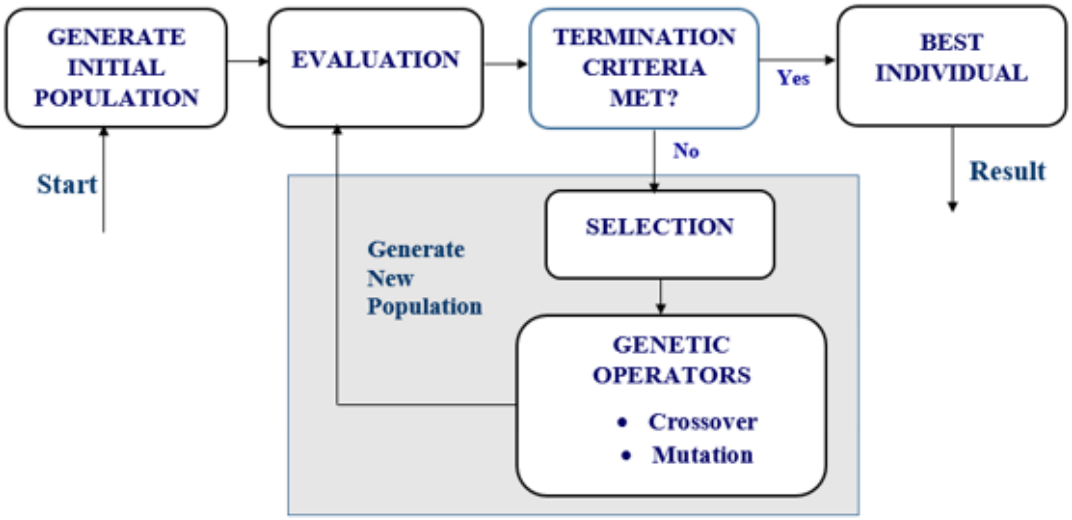


Figure 4.3: GP searching process.

Many works in GP focus on classifier induction, a task that can be accomplished by evolution using GP [280]. In GP, setting the control parameters is an important first step to manipulate data and to obtain good results. In our datasets, we tried several experiments for classification tasks by using the control parameters of GP proposed in HeuristicLab tools [281].

### 4.2.4 Performance Evaluation

The performance measures were considered based on the proportion of older people with mortality, urgent hospitalization, preventable hospitalization, disability, fracture, and access to ED with a red code. Predicting these adverse outcomes among a large number of subjects is important when applied in real-world practice. Hence, the true positive rate (TPR) was the main metric to consider. The overall accuracy (Acc), true negative rate (TNR) and F1-score, which is the harmonic mean of precision and recall, were used as additional performance metrics. The accuracy, TPR, and TNR were formulated using the true positives (TP), false positives (FP), true negatives, and false negatives (FN). These measures formally are explained in chapter 2, as defined in [47].

### 4.2.5 Data Analysis Tools

The data analysis tools used in the study are Python Scikit-learn library, RStudio software package, and HeuristicLab. In this work, the exploratory data analysis part and statistical test analysis were done using R3.5.0, whereas the entire classification problems with support vector machines, random forests, neural networks, logistic regression and decision trees were implemented using python 3.7. The Python implementation codes used in the experiment can be accessed online from the following URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7303829/. HeuristicLab is a software tool for heuristic and evolutionary algorithms, which was used, in this study, to carry out classification problems using GP.

### 4.2.6 Experimental Settings

#### Model Development

In analyzing the data for prediction, the output variables represent an occurrence in the next year, and the predictive model is proposed to predict frailty according to the expected risk of urgent hospitalization, preventive hospitalization, disability, fracture, access to ED with a red code and death within a year. The performance of various predictive models is evaluated for each outcome prediction using four metrics –Accuracy, TPR, TNR, and F1-score. These metrics provide an effective and simple way to evaluate the performance of a classifier. Using these four measures, the models were evaluated using both the holdout method and the cross-validation method. Figure 4.4 shows the general experimental workflow of the predictive machine learning model.

#### Holdout Method

In this study, our first experiment was started by exploring the predictive performance of machine learning methods using the holdout method. This method randomly splits a dataset into training and testing according to a given proportion. Each machine learning model was trained using the training dataset (70%) and evaluated using test datasets (30%). The training dataset was used for building the model, while the test dataset was used to evaluate the prediction capabilities of models.

#### K-Fold Cross-Validation

The K-fold Cross-Validation (CV) procedure was applied to each problem's

data. The CV is one of the most commonly used model evaluation procedure that extends the holdout method by repeating the splitting process several times. The K-fold CV technique divides the dataset into K folds of approximately equal size. Then, the model being evaluated is trained using the K-1 parts, and one part is left out for model validation. In this study, we used 10-folds, and the dataset was split into three parts for the purpose of model training and testing: the training set to build the model, the validation set to select the model parameters and the test set to evaluate the performance of the final model based on the selected parameters.

### Hyperparameter Tuning

In all experiments, the set of hyperparameters was selected for each ML method before the training begins. Hyperparameters allow ML algorithms to better adjust to the problem details. The hyperparameters for each model were tuned using a grid search with cross-validation in Python Scikit-learn, as described by Mueller and Guido [282]. Appendix C presents the list of hyperparameters used for training each ML model in this study.

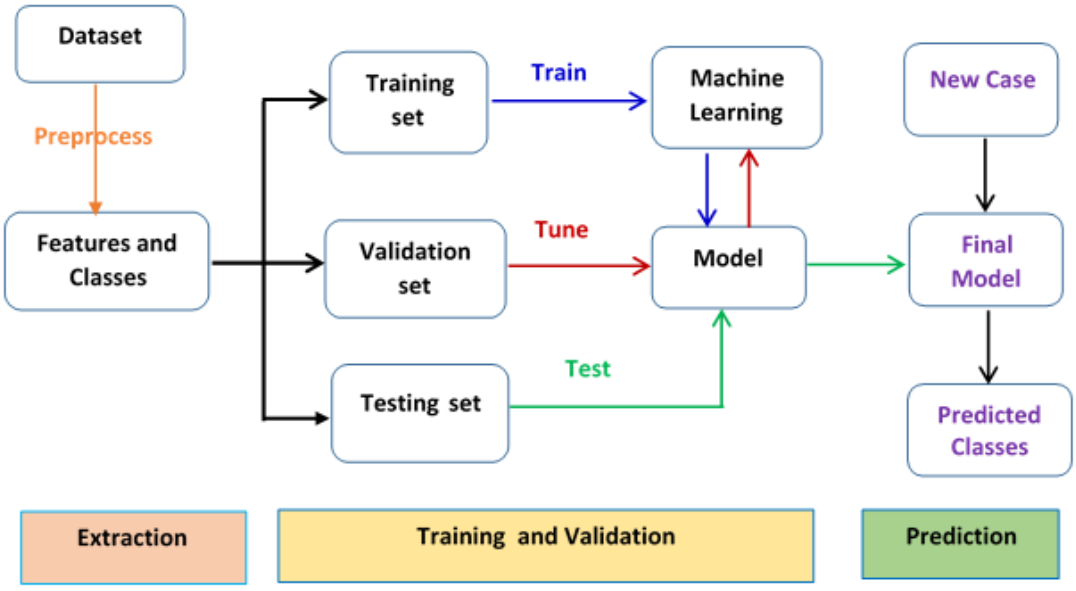


Figure 4.4: The experimental workflow of the predictive machine learning model

## 4.3 Results using Machine Learning Methods

### Study Population

From the original dataset of 1,095,612 elderly people aged 65 and above, we retrieved 83,646 with mortality, 77,836 with urgent hospitalization, 62,854 with a disability, 38,144 with preventable hospitalization, 14,978 with access to ED with red code and 14,166 with a fracture for this study. The retrieval process was made using the resampling approach, and each problem was analyzed independently of each other using the widely used machine learning models. In this section, the predictive performance of machine learning models using both holdout and cross-validation methods are presented through feature selection analysis.

### Features Selection

Feature selection provides an effective way to remove irrelevant and/or redundant features, which can reduce running time, increase learning accuracy, and facilitate a better understanding of the model. Unnecessary features can also increase the chance of overfitting and decrease the generalization performance on the test data. We used a filter method for feature selection [283]. A Chi-square test is a filter method used in this study to find out the statistical significance between features and the target. The Chi-square value, together with P-values at a significance level of .05, was used to identify the most important features with their rank, i.e., variables shown to be significantly associated with the outcome by the Chi-square test analysis (P-value $<$ .05) were selected for model building. A P-value of P$<$.001 indicates that there is an association between the input and the target variables. The strength of the association between the input variables and the target is ranked based on the Chi-square value. Out of the 58 predictor variables, 25,24,10,7,4 and 3 non-significant variables were discarded for preventable hospitalization, urgent hospitalization, emergency admission with red code, fracture, mortality, and disability, respectively. Table 4.2 presents the top 15 ranked features in order of decreasing importance in the mortality and fracture problems. The most significant feature for other problems is presented in Appendix D.

Feature importance can give us insight into a problem by indicating what variables are the most discriminating between classes. For example, in Table 4.2, 'age,' followed by the 'Charlson index,' are the most important features in the prediction of mortality, which makes sense in the problem context.

The rank of features differs from one problem to another, except for the variable 'age,' which has the highest score in all problems. Next to the 'age' attribute, variables, such as 'femur fracture,' 'the number of urgent hospitalization,' and 'neck fracture' are the most discriminant features in the fracture problem, while 'type of family' and 'home living status' are the least significant variables. 'Mental disease,' 'poly prescription' and 'disease of the circulatory system' are variables with the highest rank in the urgent hospitalization and preventable hospitalization. The 'age,' 'Charlson index,' and 'number of urgent hospitalization' are the most important predictors of emergency admission with red code.

Table 4.2: The most important variables in mortality and fracture problems

| | **Mortality problem** | | **Fracture problem** | |
|---|---|---|---|---|
| **Rank** | **Variable** | **P-Value** | **Variable** | **P-Value** |
| 1 | Age | P<.001 | Age | P<.001 |
| 2 | Charlson index | P<.001 | Femur fracture | P<.001 |
| 3 | # urgent hospitalization | P<.001 | # urgent hospitalization | P<.001 |
| 4 | # total hospitalizations | P<.001 | Neck fracture | P<.001 |
| 5 | invalidity | P<.001 | green code | P<.001 |
| 6 | # non-traumatic | P<.001 | # total hospitalizations | P<.001 |
| 7 | Disability | P<.001 | Charlson index | P<.001 |
| 8 | Poly prescriptions | P<.001 | Poly prescriptions | P<.001 |
| 9 | green code | P<.001 | invalidity | P<.001 |
| 10 | yellow code | P<.001 | Disability | P<.001 |
| 11 | blood | P<.001 | Nerve disease | P<.001 |
| 12 | Anemia | P<.001 | Depression | P<.001 |
| 13 | circulatory disease | P<.001 | blood | P<.001 |
| 14 | respiratory disease | P<.001 | Anemia | P<.001 |
| 15 | urinary tract disease | P<.001 | yellow code | P<.001 |

Some features with the lowest rank and common to urgent hospitalization and preventable hospitalization include marital status, level of education, work status, and income. Each of the predictive models (SVM, ANN, LR, RF, and DT) have been applied using the most important features in each of the six problems. GP differs from the other machine learning models

in that it performs implicit feature selection automatically during the evolutionary process. GP learns which combination of features are useful for classification and determines the optimal number of features automatically.

### Performance via Holdout Method

In this study, our first experimental results were obtained through the holdout (train-test split) method with all subsets of features using the default parameters of the models. However, these approaches have brought the problem of overfitting on the training data for RF and DT, as shown in Figures 4.5, 4.6 and 4.7. The plots indicate the accuracy without performing any parameter tuning and using all the feature subsets (from top 3-top 58 feature subsets). The left plot shows that RF and DT overfit the training data, which poorly generalizes the test data as the number of features increase. The training accuracy of DT and RF is rapidly increasing as the number of features increases while the test accuracy reduces radically with an increasing number of features. RF and DT learn the noise in the training data to the extent that it negatively impacts the performance of the model on test data.

We say overfitting occurs when a hypothesis or model doesn't generalize well from our training data to unseen data. A more formal definition of overfitting can be given as follows [71]: Given a model or hypothesis space $H$, a hypothesis $h \in H$ is said to overfit the training data if there exists some other hypothesis $h' \in H$, such that h has smaller error than h' over the training examples, but h' has a smaller error than h, over the entire distribution of examples.

In order to reduce overfitting problem and improve the performance, the parameters of each model were tuned using grid search along with the most important features associated with each outcome. Table 4.3 shows the performance of SVM, RF, ANN, and DT using the best features and parameters selected on each problem via holdout method.

In our experiments, we explored common variations for each machine learning algorithm in frailty predictions. From the results of the experiment in Table 4.3, it is clear that all algorithms behave differently for each different problem. For the mortality dataset, RF and ANN produced higher values of TPR (0.79), while the decision tree produced the lowest performance. For the fracture problem, DT has scored the highest values of TPR (0.79). The overall average TPR of RF was slightly higher for all problems, while SVM has slightly higher values of TNR in all problems, and

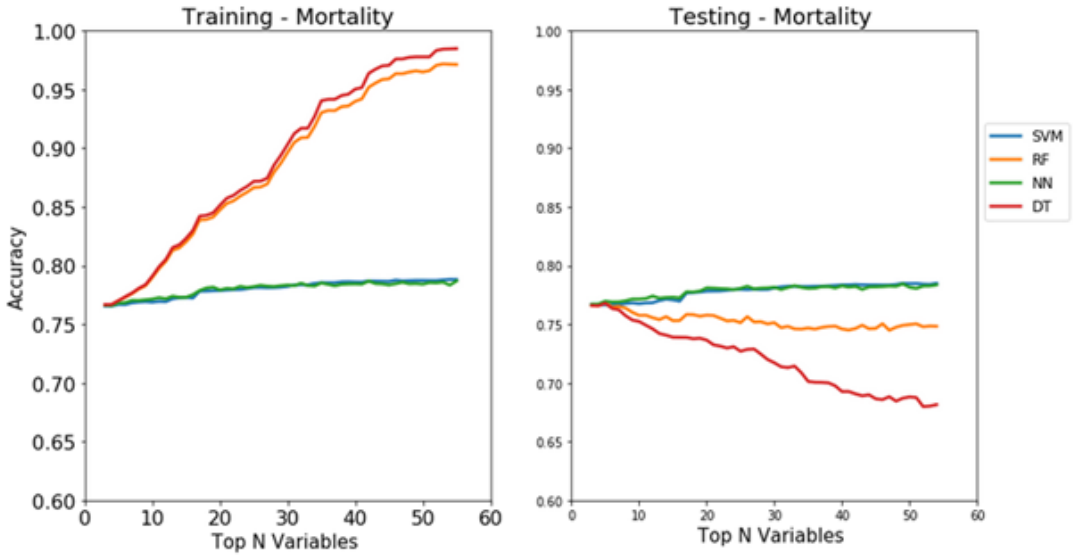


Figure 4.5: Train accuracy (left) and test accuracy (right) for mortality data with all features.

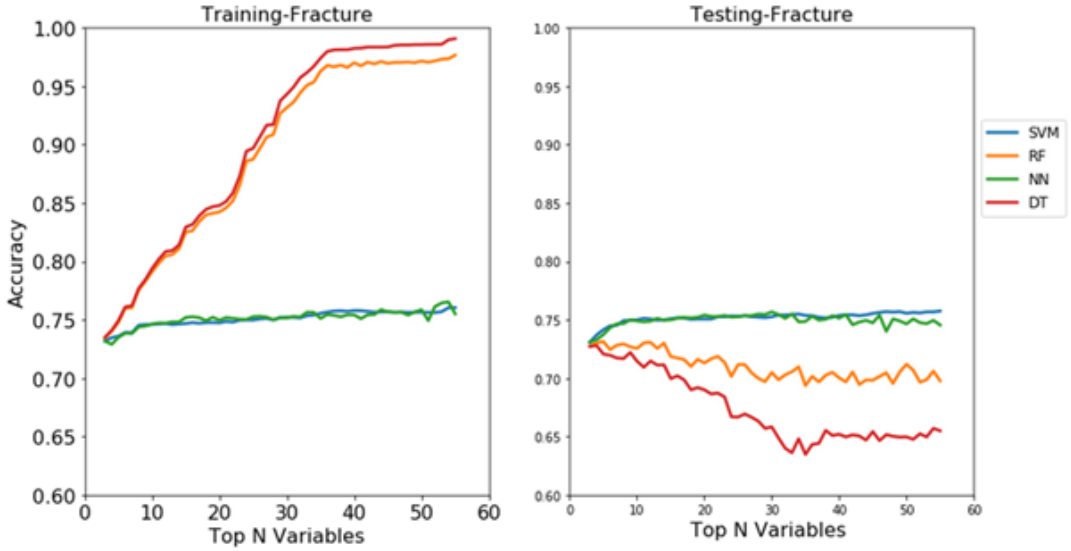


Figure 4.6: Train accuracy (left) and test accuracy (right) for fracture data with all features.

DT produced the lowest average TPR in all problems. According to the results on the test part of the dataset, all machine learning models showed lower prediction performance on the urgent hospitalization and access to ED

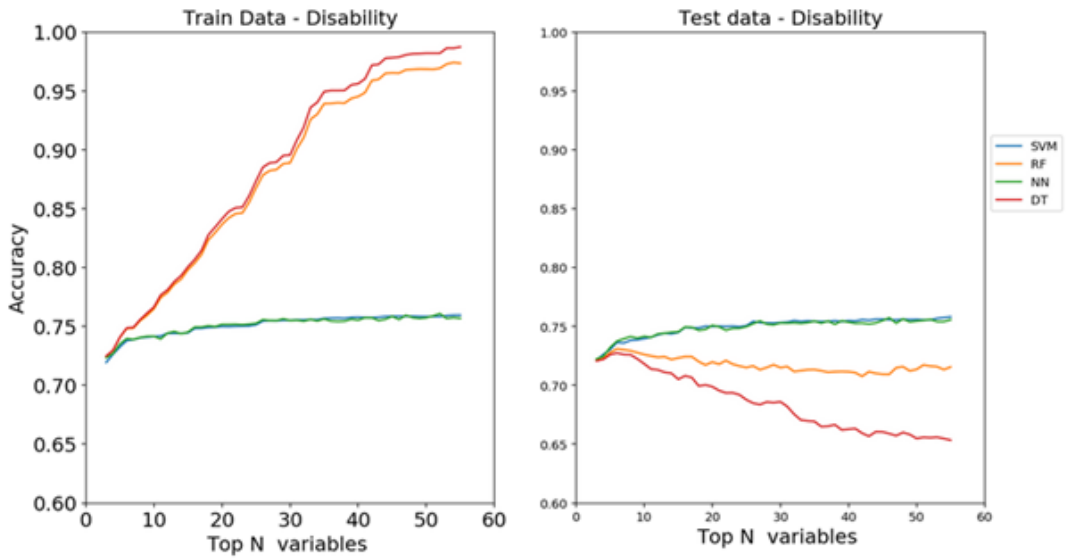


Figure 4.7: Train accuracy (left) and test accuracy (right) for disability data with all features.

with red code problems, while mortality and disability have higher values of prediction results compared to other outcomes.

Table 4.3: Prediction performance via holdout method for the six problems

| | **SVM** | | **RF** | | **ANN** | | **DT** | |
|---|---|---|---|---|---|---|---|---|
| **Problem** | **TPR** | **TNR** | **TPR** | **TNR** | **TPR** | **TNR** | **TPR** | **TNR** |
| Mortality | 0.78 | 0.78 | 0.79 | 0.77 | 0.79 | 0.78 | 0.6 | 0.79 |
| Disability | 0.78 | 0.72 | 0.78 | 0.71 | 0.75 | 0.75 | 0.78 | 0.69 |
| Fracture | 0.75 | 0.74 | 0.77 | 0.72 | 0.77 | 0.72 | 0.79 | 0.66 |
| Urgent hosptalization | 0.61 | 0.73 | 0.65 | 0.68 | 0.66 | 0.68 | 0.64 | 0.68 |
| Preventable hosptalization | 0.74 | 0.73 | 0.73 | 0.72 | 0.73 | 0.73 | 0.76 | 0.66 |
| Access ED with red code | 0.63 | 0.73 | 0.63 | 0.72 | 0.63 | 0.74 | 0.62 | 0.73 |

### Performance via 10-fold Cross-Validation

The 10-fold cross-validation reduces the variance of the resulting estimate by averaging over 10 different sub-samples. This 10-fold cross-validation

can deal with limitations of the holdout method, such as to reduce overfitting, and therefore, it is more reliable and provides better generalization performance on the test data. Thus, in our second experiment, we used 10 fold cross-validation method on each of the six datasets. The variation of each model's accuracy across the 10 samples in the 10-fold cross-validation is presented in Figures 4.8 and 4.9 for the largest dataset (i.e., mortality) and smallest dataset (i.e., fracture), respectively. From the figures, one can see that the models are more stable in predicting mortality than fracture across the 10 samples. It is also found a slight variation of classification rate across the ten samples for the other outcomes.

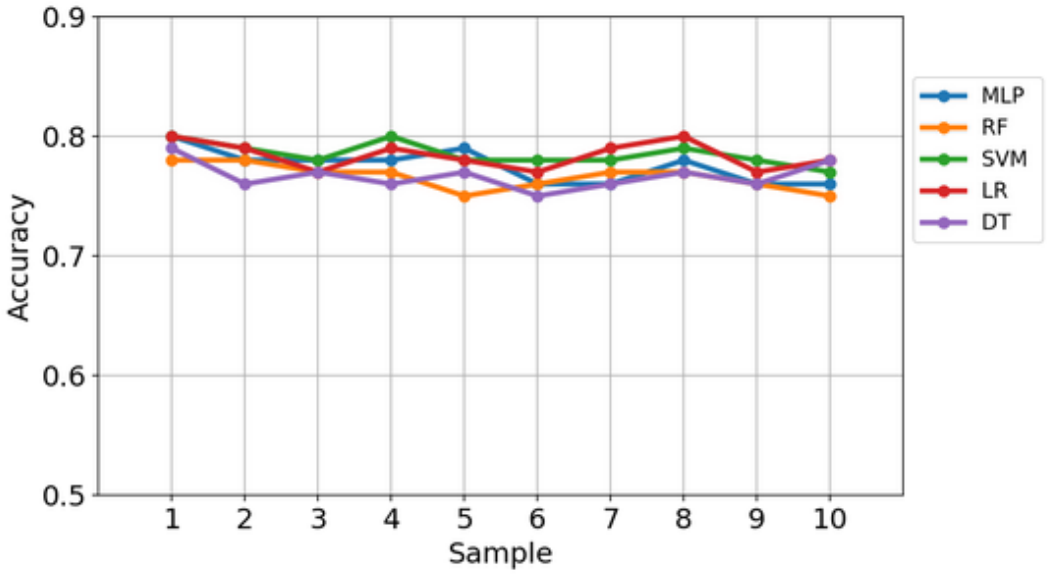


Figure 4.8: The score of five models across 10 validation samples on Mortality problem.

As shown in Figure 4.8, the classification rate across ten samples in the 10-fold cross-validation is slightly varied in each classifier for the mortality problem. The variation of accuracy is greater in the fracture problem from sample 1 to sample 10 for each model, as shown in Figure 4.9. Particularly, logistic regression has shown a greater variety of performance among the other models, where it performed the lowest accuracy at sample 7 and the highest accuracy at sample 9 in the fracture problem. The decision tree has shown the highest classification rate at sample 10 for mortality and at sample 3 in the fracture problem, while it has the lowest accuracy in the rest of the samples. The average performance of 10-fold cross-validation in each problem is shown in Table 4.4, where performance for each model is

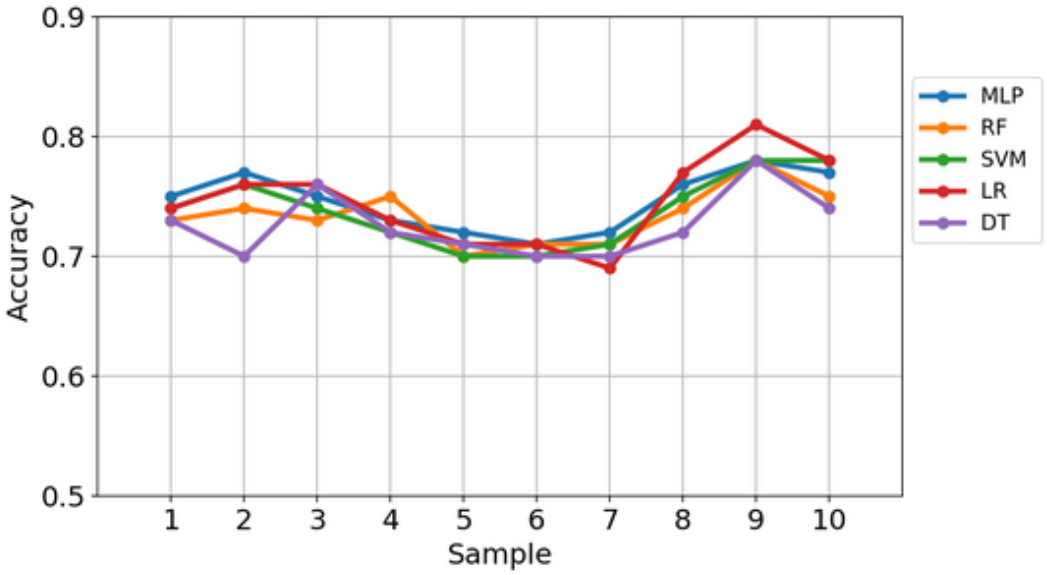


Figure 4.9: The score of five models across 10 validation samples on fracture problem.

measured using accuracy, TPR, TNR, and F1-score.

From the results of all models in each outcome presented in Tables 4.4 and 4.5, we can see that predicting mortality events has shown the highest performance, while predicting access to urgent hospitalization and access to ED with red code have shown lower performance. Next to the mortality problem, there are higher values classification rates for disability and fracture problems. This implies that the dataset in this study is better at predicting mortality than predicting the other outcomes. In predicting urgent hospitalization, only SVM achieved the best performing algorithm in all measurements (Accuracy, TPR, TNR, and F1-score) among all models trained using 10-fold cross-validation. In mortality problem, it can be seen that the highest average performance was obtained by ANN (Accuracy=0.78, TPR=0.81, TNR=0.76, F1-score=0.79) and SVM (Accuracy=0.79, TPR=0.77, TNR=0.80, F1-score=0.78) followed by LR (Accuracy=0.78, TPR=0.78, TNR=0.79, F1-score=0.78). DT produced the highest TPR (0.80) and RF showed comparable results (Accuracy=0.78, TPR=0.79, TNR=0.76, F1-score=0.76) on mortality problem. For fracture and disability problems, SVM, RF, and LR have a similar accuracy (0.75), although they all differ in TPR, TNR, and F1-score.

From the results of the experiments, it is also important to observe that the various machine learning techniques can significantly vary in terms of

Table 4.4: Prediction results of models using 10-fold cross-validation

| Models | Accuracy | TPR | TNR | F1-Score |
|---|---|---|---|---|
| **Mortality** | | | | |
| ANN | 0.78 | 0.81 | 0.76 | 0.79 |
| SVM | 0.79 | 0.77 | 0.8 | 0.78 |
| RF | 0.78 | 0.79 | 0.76 | 0.76 |
| LR | 0.78 | 0.78 | 0.79 | 0.78 |
| DT | 0.75 | 0.8 | 0.7 | 0.76 |
| **Fracture** | | | | |
| ANN | 0.75 | 0.77 | 0.73 | 0.75 |
| SVM | 0.75 | 0.77 | 0.74 | 0.75 |
| RF | 0.75 | 0.78 | 0.72 | 0.76 |
| LR | 0.75 | 0.75 | 0.75 | 0.75 |
| DT | 0.74 | 0.76 | 0.72 | 0.74 |
| **Disability** | | | | |
| ANN | 0.74 | 0.76 | 0.71 | 0.75 |
| SVM | 0.75 | 0.78 | 0.73 | 0.76 |
| RF | 0.75 | 0.77 | 0.72 | 0.75 |
| LR | 0.75 | 0.76 | 0.73 | 0.74 |
| DT | 0.73 | 0.78 | 0.7 | 0.75 |

Table 4.5: Prediction results of models using 10-fold cross-validation

| **Models** | **Accuracy** | **TPR** | **TNR** | **F1-Score** |
|---|---|---|---|---|
| **Urgent hospitalization** | | | | |
| ANN | 0.67 | 0.6 | 0.7 | 0.66 |
| SVM | 0.75 | 0.8 | 0.7 | 0.76 |
| RF | 0.66 | 0.7 | 0.7 | 0.66 |
| LR | 0.67 | 0.7 | 0.6 | 0.65 |
| DT | 0.66 | 0.7 | 0.7 | 0.65 |
| **Preventable Hospitalization** | | | | |
| ANN | 0.74 | 0.7 | 0.7 | 0.73 |
| SVM | 0.74 | 0.7 | 0.8 | 0.73 |
| RF | 0.73 | 0.7 | 0.7 | 0.73 |
| LR | 0.74 | 0.7 | 0.8 | 0.73 |
| DT | 0.72 | 0.7 | 0.7 | 0.72 |
| **Access to ED with red code** | | | | |
| ANN | 0.7 | 0.7 | 0.7 | 0.67 |
| SVM | 0.68 | 0.6 | 0.7 | 0.66 |
| RF | 0.68 | 0.7 | 0.7 | 0.67 |
| LR | 0.69 | 0.6 | 0.7 | 0.67 |
| DT | 0.67 | 0.7 | 0.7 | 0.68 |

their performance for the different evaluation metrics. For example, in the mortality problem, SVM outperforms DT and ANN in TNR value (0.80), and ANN outperforms both SVM and DT in F1-score (0.79), while DT outperforms both models in TPR value (0.80). It also noted that the performance of all models differs in all problems due to the difference in feature space, size, and diversity of data in each of the six problems. This can be seen that the prediction performance of all models trained with mortality data (largest in size) is much better than the performance of models trained with access to ED with red code data (smaller in size), which demonstrates that the size of data is an important factor for better performance, but not always true for all models. In addition, the performance of each ML technique varies from problem to problem. For example, the performance of ANN measured in TPR is 0.81, 0.77, 0.76, 0.74, 0.70 and 0.67 for mortality, fracture, disability, preventable hospitalization, access to ED with red code and urgent hospitalization, while for DT the TPR is 0.80,0.75,0.78, 0.73, 0.70 and 0.65 for each problem, respectively. Considering the performance of these two ML methods (ANN and DT) in their TPR value, ANN outperforms DT in mortality and fracture problems, while DT outperforms ANN in the disability and access to ED with red code problems. We can also see that LR has a higher TPR value than SVM in the mortality problem. This shows that it is not necessarily that the more complex ML models (e.g., ANN, SVM) can always outperform simpler models (e.g., DT, LR). The RF classifiers are considered to be less complex than SVM and ANN. In 10-fold cross-validation, however, it achieved comparable performance to SVM and ANN in most of the problems. On the other hand, tree-based classifiers (RF and DT) are more sensitive to bad features and quality of data. Therefore, effective feature selection is an important step to improve their performance. SVM model tends to perform well in high-dimensional classification problems; however, it may not perform well if the sample classes of the problem are highly overlapping. ANN can generally outperform other techniques if the dataset is very large and if the structure of the dataset is complex (e.g., if they have many layers).

In general, machine learning is an exploratory process, where there is no one-size-fits-all problem. In particular, there is no model that is recognized to achieve supreme performance for all problem types, domains, or datasets. The best performing ML model differs from one problem to another according to the characteristics of variables, the size of the data, and the metrics used. The idea is similar to the "No Free Lunch" theorem [284], which

states that there is no universal algorithm that works best for every problem. However, it is important to study each problem by evaluating each model carefully in order to reach an effective predictive design. The results also show that it is essential to carefully explore and evaluate the performance of ML techniques using various optimized parameter values as well as using the most significant predictor variables. Particularly, tree-based classifiers (e.g., RF and DT) are more sensitive to overfitting problems, as shown in Figures 4.5 - 4.7 on the mortality, fracture and disability problems, if the correct subset of features is not selected or if the required parameter values of models are not configured properly. The accuracy in the figures clearly indicates that an increasing number of features in RF and DT leads to the model overfitting. Interestingly, SVM and ANN models showed relatively consistent performance both on training and testing, even with an increasing number of features.

## 4.4 Results using Genetic Programming

In this section, we investigated the performance of GP for the prediction of frailty status in terms of the six problems or outcomes. The predictors common to all problems and which were also included in the final model produced by GP were age, the number of urgent hospitalization, Charlson comorbidity index, dementia, and mental disease. The final prediction model of each problem generated by GP is a binary parse tree representing the classification model.

### GP Parameter setup

In GP, setting the control parameters is an important first step to manipulate data and to obtain good results. In our datasets, we tried several experiments for classification tasks by using the control parameters of GP, such as population size, selection method, number of elite individuals, initialization method, number of generations, crossover probability rates, and mutation probability rates. Due to the stochastic nature of GP, 30 runs were performed in all problems, each with a different random number generator seed. For our frailty problem, we specifically focused on the two common parameters of GP: Maximum number of generations and population size. In order to investigate the effect of few generations over a larger population and small population over more generations and also to get an advantage from either of these GP parameter settings, we run two different algorithms

of GP (GP1 and GP2) under varying population size and the maximum number of generations, keeping all other parameters set to default. The maximum number of generations and population size for GP1 is set to be 1000 and 100, respectively. In GP2, we set a maximum number of generations to be 100 and population size 1000. For all frailty problems, GP1 and GP2 were applied, and for each experiment, 30 runs were performed with the same initial configurations of parameters. We clearly observed that the runs with a population size of 1000 and generation 100 are related to the immense runtime requirements, comparing with the runs of population size 100 and generation 1000. In fitness, it is apparent that a large population running for a small number of generations behaves differently from the small population running for a large number of generations. The summary of parameters used for running GP2 experiments is presented in Table 4.6.

Table 4.6: GP Control Parameters used in the Experiment

| **Parameter Name** | **Value** |
|---|---|
| Algorithm | GP2 |
| Maximum generations | 100 |
| Population size | 1000 |
| Mutation rate | 15% |
| Crossover rate | 90% |
| Solution creator | Ramped half and half |
| Max tree depth | 10 |
| Max tree length | 100 |
| Elites | 1 |
| Terminal set | Constants, Variables , |

The fitness of GP1 and GP2 across generations were compared for mortality and fracture problems using mean squared error (MSE). The MSE is used as fitness to compare the quality of the two models (GP1 and GP2), and it was observed that GP2 produced lower error rates, which is ranging from 0.18 to 0.25 for mortality and from 0.19 to 0.25 for fracture problems. While for GP1, the MSE is much higher, which is ranging from 0.20 to 0.30 for the mortality and from 0.22 to 0.29 for fracture problems. The results show that a large population is more likely than a small population to make more significant improvements in fitness from one generation to the next,

given that it generates more new trees in each generation. Generally, for frailty problems, it seems that results with GP2 are more stable and that larger population is a better choice than many generations. As a result of this, we preferred GP with a larger population size and a smaller number of generations for the prediction of frailty conditions.

### GP Prediction Performance

In analyzing GP for classification, the most important aspect is to know the number of samples that are classified correctly and those, which are classified incorrectly. The results averaged from 30 runs of GP experiments are presented in Table 4.7 on the training set, and Table 4.8 on the testing set.

Table 4.7: Performance of GP on the training set

| Problem | Sensitivity (SD) | Specificity (SD) | Accuracy (SD) |
|---|---|---|---|
| Mortality | 0.75(0.05) | 0.75(0.06) | 0.75(0.02) |
| Access to ED with red code | 0.76(0.24) | 0.45(0.37) | 0.58(0.09) |
| Disability | 0.72(0.04) | 0.69(0.05) | 0.72(0.02) |
| Fracture | 0.71(0.04) | 0.67(0.14) | 0.74(0.08) |
| Urgent hospitalization | 0.65(0.22) | 0.63(0.29) | 0.64(0.13) |
| Preventable hospitalization | 0.71(0.18) | 0.63(0.33) | 0.67(0.11) |

Table 4.8: Performance of GP on the testing set

| Problem | Sensitivity (SD) | Specificity (SD) | Accuracy (SD) |
|---|---|---|---|
| Mortality | 0.75(0.05) | 0.76(0.06) | 0.75(0.02) |
| Access to ED with red code | 0.73(0.24) | 0.43(0.36) | 0.58(0.08) |
| Disability | 0.70(0.04) | 0.73(0.05) | 0.71(002) |
| Fracture | 0.71(0.14) | 0.67(0.08) | 0.72(0.04) |
| Urgent hospitalization | 0.66(0.22) | 0.62(0.29) | 0.63(0.13) |
| Preventable hospitalization | 0.73(0.18) | 0.64(0.33) | 0.68(0.11) |

In these problems, using sensitivity and specificity allows us to correctly identify those with the disease condition (frail people) and to correctly identify those without the disease (non-frail people), respectively. The standard

deviation (SD) for mean sensitivity, specificity, and accuracy are also calculated since each problem is run 30 times, as shown in Tables 4.7 and 4.8. For the mortality problem, GP produced the best performance in all measurements. For access to ED with red code, the overall accuracy and specificity of GP are slightly lowered. For the remaining problems, the performance of GP is at an acceptable level. These results confirmed the predictive capability of GP on frailty problems.

### Comparison of GP with other non-GP classifiers

In the literature, there are some studies that compare GP with other statistical and machine learning methods [285]. The studies suggest that GP may be better at representing the potentially non-linear relationship of (a smaller subset of) the strongest predictors, although the complexity of the GP-derived model was found to be much higher. The fact that GP requires fewer variables to achieve similar performance may have an advantage in the practical application of the developed clinical prediction models. Therefore, a prediction model that requires fewer inputs, especially if the information relating to these inputs is in practice recorded easily and to good quality, would considerably increase adoption and utility. A comparison of GP with statistical models, such as cox regression techniques [286], was attempted in terms of the performance of a cardiovascular risk score using a prospective cohort study of patients with symptomatic cardiovascular disease. The predictive ability of the Cox regression model and GP was evaluated in terms of their risk discrimination and calibration using the validation set. Their findings indicated that the discrimination of both models was comparable. Using the calibration of these models, which was assessed based on calibration plots and the generalization of the Hosmer-Lemeshow test statistic, was also similar, but with the Cox model is better calibrated to the validation data. In [287], a comparison of GP and ANN in metamodeling of discrete-event simulation was studied. The results of this study concluded that GP provides greater accuracy in validation tests, demonstrating a better generalization capability than ANN, despite the fact that GP, when compared to ANN, requires more computation in model development. Most machine learning methods are usually straightforward to implement and work well with minimum resources; however, their black-box nature makes them non-user-friendly. On the other hand, GP results are often human friendly and provide an explicit mathematical formula as its output, although developing such an efficient algorithm and realizing its full potential to solve real-world

problems can be challenging. GP algorithms are expected to require a computing time that grows exponentially with the size of the problem [288]. Most commonly, the ability of a machine learning algorithm to produce high performance results depends on the nature of the problem as there is no single algorithm that works best for every problem. As a result, we compared the results of GP with the other commonly used machine learning models in terms of prediction performance on the six different problems of frailty: mortality, access to the emergency department with red code, disability, fracture, urgent hospitalization, and preventable hospitalization. On each of the six problems, the results of GP were compared with support vector machine (SVM), random forest (RF), artificial neural network (ANN) and decision tree (DT).

In each of the six problems associated with frailty, the results obtained from the non-GP classifiers (SVM, RF, ANN, and DT) are compared with the results of GP using sensitivity. The comparison is based on the ability to identify the positive subjects in the frailty problems using their respective datasets. The performance comparison of predictions on four problems by the different classifiers is shown in Figures 4.10 and 4.11. The figures depict the performance of all classifiers using sensitivity on the testing part of the data. From the figures, the performance values were obtained using a different subset of ranked features, the boxplots represent the performance at every 30 runs of GP, and the different colored dots represent the performance of the other machine learning algorithms. In all plots, the x-axis represents the number of features, and the y-axis represents the performance of GP using sensitivity.

Looking at each box plot of GP in Figures 4.10 and 4.11, we can observe that some runs are outliers in each problem due to the stochastic nature of GP. For example, in urgent hospitalization, there are three runs beyond the whiskers for the top 5 and top 10 variables. These runs are outliers of the 30 runs of GP, plotted as points. In all problems with all variables, the performance of SVM, RF, ANN, and DT are displayed under the upper quartile of the GP box plots, indicating the maximum performance obtained from the 30 runs of GP is always greater than the performance of the machine learning models. Comparing all algorithms, the decision tree followed by random forest has the lowest performance in all problems for the number of variables greater 10. The average sensitivity of GP overlaps with the performance of ANN. However, the accuracy of GP is lowered compared to

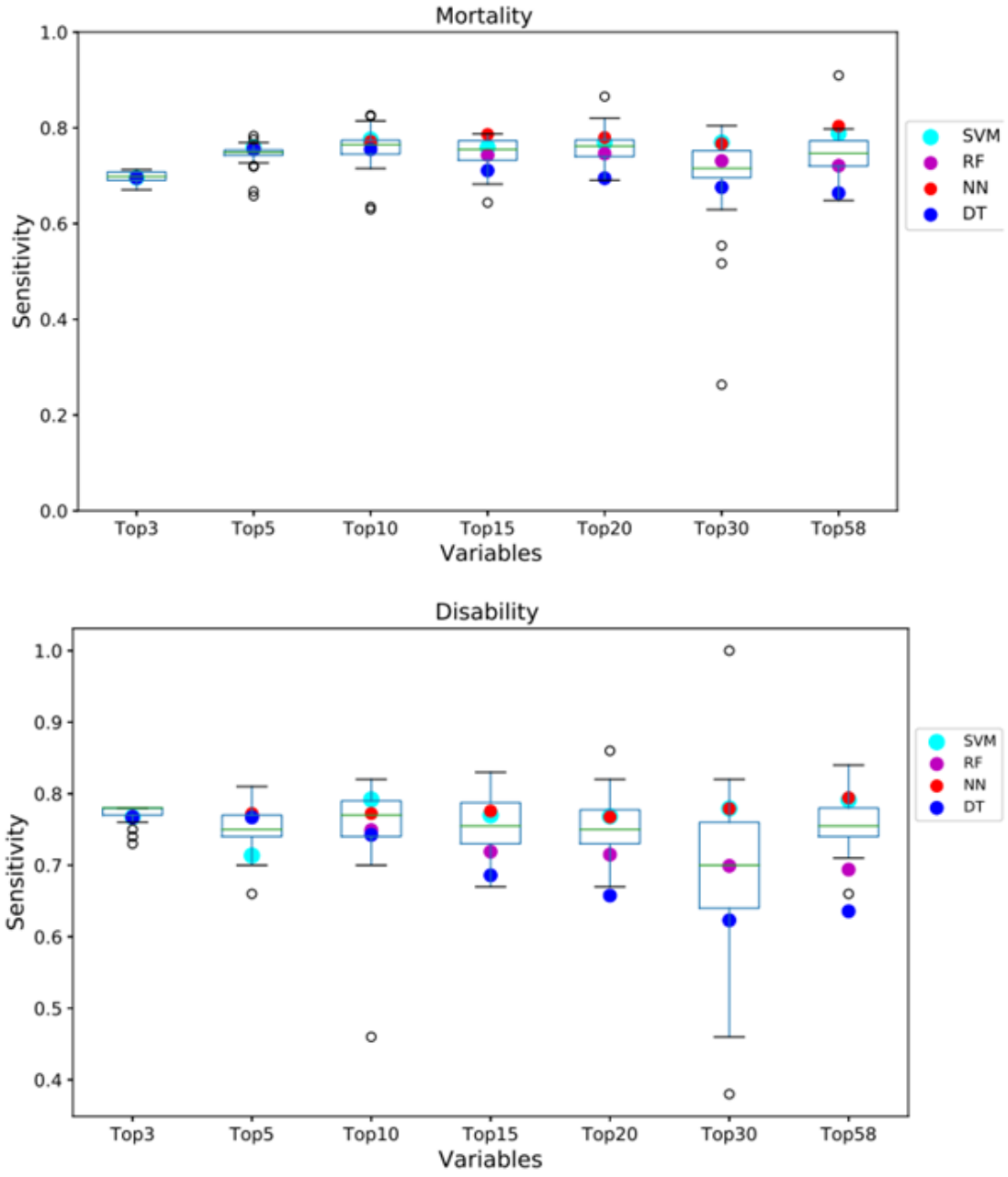


Figure 4.10: Performance of GP on Mortality (upper plot) and Disability (lower plot) problems compared to other algorithms.

SVM and ANN. For making the fairest comparison possible between GP and other machine learning models, a pairwise statistical test between the 30 runs of GP and each individual machine learning model was also performed. The statistical test used was the Wilcoxon signed-rank test. The

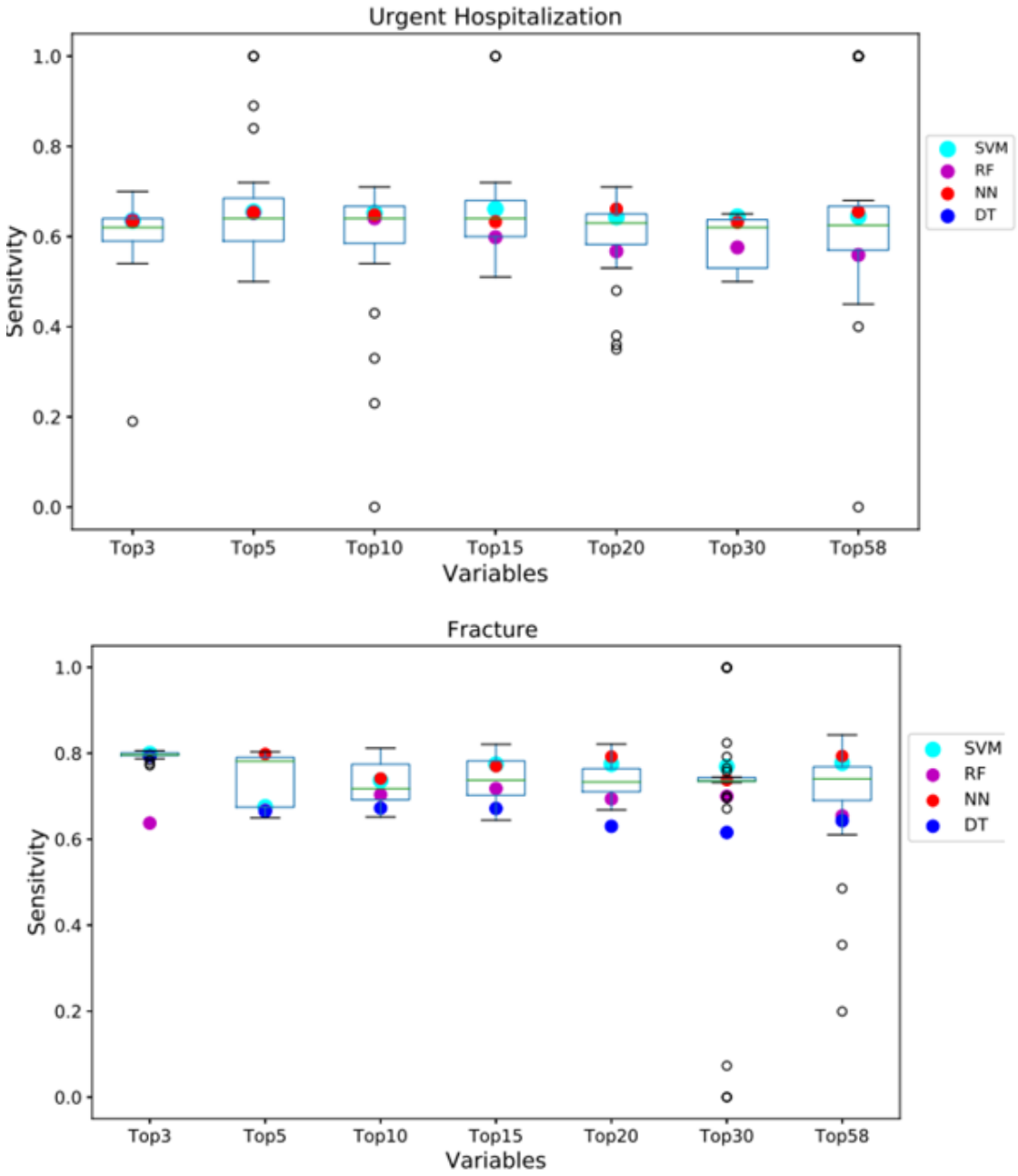


Figure 4.11: Performance of GP on Urgent hospitalization (upper plot) and Fracture (lower plot) problems compared to other algorithms

Wilcoxon statistical test is a nonparametric test that ranks the differences in performances of GP and other algorithms over each frailty problem. The test was based on the sensitivity score of each algorithm in each problem

on the test data at the significance level of 0.01. The results of the test in terms of P-values with the significance level of 0.01 are shown in Table 4.9.

Table 4.9: Results of Wilcoxon signed-rank test in terms of P-values

| Problem/dataset | SVM vs. GP | RF vs. GP | NN vs. GP | DT vs. GP |
|---|---|---|---|---|
| Mortality | 0.000 | 0.003 | 0.001 | 0.000 |
| Fracture | 0.000 | 0.021 | 0.000 | 0.002 |
| Disability | 0.059 | 0.004 | 0.012 | 0.003 |
| Urgent hospitalization | 0.709 | 0.013 | 0.374 | 0.013 |
| Preventable hospitalization | 0.682 | 0.026 | 0.871 | 0.005 |
| Access to ED with a red code | 0.006 | 0.000 | 0.011 | 0.000 |

As depicted in Table 4.9, the Wilcoxon test allows rejecting 11 hypotheses. The P-values below 0.01 indicate that the respective algorithms differ significantly in TPR, while the P-values above 0.01 indicate that the algorithms behave similarly in predicting frailty conditions. The test results between SM and GP are statistically significant only in disability, urgent hospitalization, and preventable hospitalization. Combining the experimental results and Wilcoxon signed-rank test results, it is concluded that for mortality and fracture SVM outperformed GP in TPR score, while GP outperformed SVM and RF on urgent hospitalization and access to ED with red code. Despite the fact that DT represented higher values of TPR on the preventable hospitalization compared to other algorithms, its lowest TNR result represented a higher disadvantage. ANN has a similar performance with GP for preventable and urgent hospitalization events.

### Feature Selection Comparison of GP and Chi-Square

The performance of GP feature selection is compared with the well-known Chi-square feature selection method. The top three variables (age, Charlson index, and the number of urgent hospitalization) selected by GP are also selected by Chi-square as the top three variables in the mortality problem. After three variables, there is slightly a little difference in the position of variables. Table 4.10 presents the prediction accuracy of the classification model using the features selected by GP and Chi-square for all problems.

For each problem, the best average accuracy of the 30 runs of GP is taken to compare the classification performance of GP and Chi-square feature selection methods. From this table, Chi-square performed the best in the mortality problem with an accuracy of 76%, followed by GP with an accuracy of 75%, a difference of only 1%. This condition also holds for disability and fracture problems. For urgent hospitalization, both GP and Chi-square produce a similar performance. The results show that GP can perform the feature selection task with competitive results.

Table 4.10: Prediction accuracy via feature selection of GP and Chi-square

| Problem | GP feature selection | Chi-square feature selection |
|---|---|---|
| Mortality | 0.75 | 0.76 |
| Urgent hospitalization | 0.64 | 0.64 |
| Disability | 0.72 | 0.73 |
| Preventable hospitalization | 0.68 | 0.71 |
| Access to ED with red code | 0.58 | 0.68 |
| Fracture | 0.71 | 0.73 |

## 4.5 Discussions

### Principal Results

The goals of the study under this chapter were to develop models to predict the risk of hospitalization, disability, mortality, fracture, and emergency admissions among older people in Piedmont, Italy. We inspected the possibility of using an administrative dataset to detect frailty in older adults using different machine learning models, which have been used as a potential tool for developing a prediction model. Six different models were developed, and the performance of each model relies on the input data provided to the learning algorithm. The performances of models were assessed by splitting the data into a training set and test set. The test set was untouched during the entire training and model selection process and only used for the final model evaluation. A predictive model that can use administrative health data will be useful in various settings to classify those individuals who are at risk of frailty and to deliver preventive interventions. Several experiments were conducted using different classification techniques to build predictive

models for frailty. The results show that various ML models can vary significantly from problem to problem in terms of the different evaluation metrics. The explored models have shown solid predictive power to estimate the risk of mortality than predicting disability, fracture, emergency admissions in red code, urgent hospitalization, and preventable hospitalization within the next year. Although each model is not a comprehensive model to predict all frailty outcomes, we have demonstrated that the SVM model has shown higher overall accuracy (0.79) in predicting mortality and urgent hospitalization than other models, when using 10-fold cross-validation. On the other hand, except for the ANN, all other ML models have shown relatively poor overall accuracy in predicting emergency admission with red code.

In addition, our results show significant performance enhancement by reducing features. In order to reduce the overfitting problem and to improve the prediction performance of classifiers, the feature selection process is executed, where the best subset of the available features is chosen. In each binary classification problem, all the independent variables were ranked using the Chi-square feature selection method for each outcome in both holdout and cross-validation methods. Using 10-fold cross-validation on mortality problems, the TPR values (also called, sensitivity) of ANN, SVM, RF, LR and DT were 0.81, 0.77, 0.79, 0.78 and 0.80, respectively. In the holdout method, almost similar results have been obtained for ANN, SVM, and RF, while DT produced higher TPR values using 10-fold cross-validation than holdout method on the mortality problem. In general, 10-fold cross-validation reduces variance by averaging over 10 different partitions, and then, it is less sensitive to any of the partitioning bias in the training and testing data. On predicting emergency admission with red code, GP achieved better TPR value than SVM, ANN, LR, RF, and DT, while SVM outperformed all models in predicting urgent hospitalization in all evaluation measures.

Generally, an important observation from the results of the experiments is that, on average, some of the ML models produce quite similar results from the same outcome, while the best performing model varies from one outcome to another outcome in terms of different metrics. For example, SVM and ANN produce similar average performance across all evaluation metrics in mortality and hospitalization outcomes. RF and LR produced similar performance on average across all measurements in disability and fracture outcomes. However, the prediction results of each ML model varies from mortality to fracture or fracture to hospitalization, etc. This can

demonstrate the feasibility of identifying frail older subjects through routinely collected administrative health databases.

**Strength and limitation**

The strength of our study is the possibility to include a multidimensional administrative database using the most powerful predictive machine learning models. In contrast to the previous studies, the prediction models use a wide variety of input variables, including clinical and socioeconomic aspects, with six simultaneous outcomes. The use of routinely collected socio-clinical data can represent the multidimensional loss of an individual's reserves, which allows predicting prospective outcomes in the elderly. Moreover, the predictions of frailty in terms of the six different adverse outcomes were assessed and analyzed, which is a step forward in studying the association of frailty with multiple health conditions on a frail person. There are limitations to our study. Despite the original data comes with multiple outcomes, each machine learning algorithms were designed to predict a single outcome, and each result is analyzed independently of each other. Therefore, further studies can be investigated to construct a predictive model that considers the correlations among the output variables to provide a list of relevant outputs for a given, previously unseen patient. Furthermore, the patients' information such as gender can be included in the study in order to understand gender-related factors for frailty and their impact on hospitalization and mortality among older people.

## 4.6 Conclusions

Predictive modeling using the information available from the administrative health database is an efficient method to identify frail older people appropriate for interventions to prevent adverse outcomes. The proposed predictive models can be applied to detect and predict frail people who are at increased risk of adverse outcomes. This study suggests that a machine learning-based predictive model could be used to screen future frailty conditions using clinical and socioeconomic variables, which are commonly collected in community healthcare institutions. With efforts to enhance predictive performance, such a machine learning-based approach can further contribute to the improvement of frailty interventions in the elderly community.

# Chapter 5

# Multi-label Classification for Early Diagnosis of Frailty Syndrome

In this chapter, we study a multi-label classification problem for the simultaneous prediction of multiple outcomes of frailty conditions: mortality, fracture, disability, medical emergency admission at the emergency department, urgent hospitalization, and preventable hospitalization. The models are constructed based on state-of-the-art multi-label classification algorithms, including binary relevance, classifier chains, and label powerset approaches. In addition, a hybrid of SMOTE and Tomek links (ML-TLSMOTE) is proposed to address the inherent problem of an imbalanced dataset in multi-label classification. More specifically, the proposed hybrid approach is designed to solve the joint occurrence of minority and majority labels in the same data patterns in multi-label learning. The proposal was tested using the imbalanced multi-label dataset of older adults aged 65 years and above. The experimental results show that the proposed method works effectively on the imbalanced label distributions. Moreover, the multi-label modeling of frailty helps to investigate label correlations and provide insights to tackle future complications associated with frailty.

## 5.1 Introduction

Frailty syndrome is highly prevalent among older people and is mostly associated with multimorbidity [289], which leads to several adverse health

outcomes. It is also a real challenge for many societies from social, financial, and economic perspectives. It is commonly recognized that frailty aggravates the risk of poor outcomes (e.g., hospitalization, loss of autonomy, functional impairment, and death) and that it escalates health and social challenges [290]. There are several tools that have been used for the detection of frailty. The Fried Phenotypic Model [20] is one of the most widely used tools for assessing physical frailty. It is based on the assessment of five measurable components: slow walking speed, self-reported exhaustion, low grip strength, unintended weight loss, and low physical activity. According to this model, frailty is considered to be present if a person has at least three of the above pre-defined components. Following the concept of the Fried model, several studies have been conducted to estimate the prevalence of frailty in older people [291].

However, it should be noted that there is still considerable uncertainty around the concept of frailty [13] and that a phenotypic evaluation of subjects is challenging when considering a large population. There are many reasons as to why it is so difficult to define and conceptualize frailty, including its complex etiology [292], the often independent work of researchers in diverse areas of frailty, such as biological basis, social basis, environment, and technology, etc. [293, 294], and the inherent difficulty in distinguishing frailty from aging and disability [295]. There is also a considerable degree of heterogeneity among the different studies of frailty models in terms of sample type and size, population characteristics and settings, baseline frailty status, and outcomes. In general, the current challenges of frailty research include the lack of standard definition of frailty which leads to lack of a standard screening and diagnostic tool, further understanding of interventions to reverse frailty, the best time for intervention, comprehensive and common understanding model to face the challenges and early estimation of multiple adverse outcomes in a frail patient [296, 297].

Until now, the state-of-the-art statistical or analytical considerations have been targeted for the intervention of a single outcome or risk factor associated with frailty. For example, the Fried's frailty phenotype was specified as a significant risk factor for 6-month adjusted mortality but was not associated with delirium and in-hospital falls [40]; similarly, a frailty risk model developed by Bertini et al.[263] predicts all-cause mortality within a year, another study proposed in [298] was designed to predict hospital admissions of older persons based on healthcare data. Our recently published work on frailty [302] has also been focused on single outcome prediction,

where separate models were developed for predicting mortality, hospitalization, fracture, and disability. Clinically, however, it makes sense for the interventions to target more than one simultaneous outcomes with common heterogeneous risk factors associated with frailty. This is due to the fact that the co-existence of multiple chronic conditions or comorbidity is common in older people [299, 300], which contributes to having multiple adverse outcomes. Therefore, this study aims to construct a predictive model that considers the correlation among multiple outcomes to provide a list of relevant outputs for a previously unseen patient. In other words, we developed a multi-label classification (MLC) model to predict the six outcomes of frailty simultaneously: mortality, medical emergency admission at an emergency department, urgent hospitalization, disability, fracture, and preventable hospitalization.

MLC is focused on training prediction functions that can associate an instance with multiple labels that are not necessarily mutually exclusive [301]. These days, MLC has gained considerable attention in the machine learning community, which appears in many application domains, and it is natural for many real-world problems, such as clinical diagnosis, disease prediction, activity recognition, object detection, image classification, etc. The existing methods for the MLC task are problem transformation and algorithm adaptation methods. The former transforms the MLC task into one or more single-label classification [176], regression problems, or label ranking [174] tasks, while the latter could extend specific learning algorithms to handle multi-label dataset directly[175].

One of the main challenges of the MLC problem is the existence of imbalanced labels where the number of 1's (positive class) in one label is much larger than the number of 1's in the other label. Such imbalanced label distributions are the intrinsic characteristics of most multi-label datasets. More specifically, the joint occurrence of minority and majority labels in the same instances affects the prediction performance of multi-label learning methods. In order to deal with this problem, we proposed a hybrid of SMOTE and Tomek links (ML-TLSMOTE) to reduce the imbalanced label distributions while diagnosing frailty conditions. Until now, such a hybrid approach has only been applied to single-label classification. In this paper, it has been extended to the multi-label scenario as an alternative solution for addressing the imbalanced problem.

## 5.2 Methods and Materials

### 5.2.1 Data Source

Data were already described in section 4.2.1 of chapter 4. Briefly, to develop a multi-label predictive model, we used health information retrieved from two years of administrative databases of elderly peopled age 65 years and above. Data are collected using an individual record linkage between the Italian 2011 census and the administrative health databases (enrolees' registry, hospital discharges, drug prescriptions, outpatient clinical investigation database, and health exemptions).There are 1095613 anonymous record items consisting of input variables such as demographic, socioeconomic, and chronic conditions and output variables, which are described as outcomes or measurable changes in the health status of patients. In this study, six output variables that are associated with each individual's status are used as labels. They are mortality, urgent hospitalization, medical emergency admission at emergency department, disability, fracture, and preventable hospitalization.

### 5.2.2 Data Description

All the six labels (i.e., the outcomes) in the data are binary-valued, as shown in Table 5.1, which presents randomly selected records from the original dataset. Labels that are associated with each record are called relevant (or

Table 5.1: An example of multi-label data records with six labels

| Records | Label 1 | Label 2 | Label 3 | Label 4 | Label 5 | Label 6 |
|---|---|---|---|---|---|---|
| $r_1$ | 0 | 0 | 1 | 0 | 1 | 1 |
| $r_2$ | 0 | 0 | 1 | 1 | 1 | 0 |
| $r_3$ | 0 | 1 | 1 | 1 | 0 | 0 |
| . . . | 1 | 1 | 1 | 0 | 0 | 1 |
| $r_m$ | 1 | 1 | 1 | 1 | 1 | 0 |

active) labels, whereas the remaining (i.e., the non-associated labels) are the irrelevant ones. For example, in Table 5.1, labels 3, 5, and 6 are relevant (associated) to the first record, while labels 1, 2, and 4 are the irrelevant ones (non-associated labels).

We used label cardinality (Card) and label density (Dens) to describe the

characteristics of our dataset. Label cardinality of dataset M, denoted by card (M), is the average number of labels of examples in M. Label density of dataset M, denoted by dens (M), is the average number of labels of examples in M divided by the number of labels. These measures are defined in section 3.3.4 of chapter 3. Table 5.2 shows the summary of the original dataset in terms of Card, Dens, number of input features (NF), the number of labels $q$, and the number of distinct label combinations (DC).

Table 5.2: Description of the multi-label dataset in the experiment

| Dataset | Instances | NF | \|L\| | DC | Card | Dens |
|---|---|---|---|---|---|---|
| Frailty | 1,095,613 | 58 | 6 | 64 | 0.13 | 0.02 |

### 5.2.3 Multi-Label Classification

Multi-label classification (MLC) problem is a generalization of a single label (binary or multi-class) classification problem where an instance is associated with more than one label simultaneously. In this study, the frailty risk prediction problem is formulated into a multi-label classification problem. Given a set of $m$ medical records $M = \{r_1, r_2, ..., r_m\}$ and a finite set of $q$ outcomes $L = \{\lambda_1, \lambda_2, ..., \lambda_q\}$, each record in M is associated with one or more outcomes in L. In this context, the 'outcomes' represent the labels. The set of multi-label training examples of the frailty classification problem can be represented by $S = \{(r_i, Y_i), i = 1, ..., m\}$, where $r_i$ is the feature vector and $Y_i \subseteq L$ denotes the set of labels for the $i^{th}$ record. The objective is to build a classification model to predict a set of labels $Y_i'$ for every new record $r_i'$. In this study, for any patient, multiple outcomes were identified in the data, and each outcome is considered as a label.

There are several multi-label classifiers to train on a multi-label dataset (MLD) [96]. The most common and most straightforward approach is binary relevance (BR). It considers each label as an independent problem and trains one binary classifier per label. BR is the baseline MLC algorithm that does not consider the relationships that may exist between labels. To overcome this limitation, several ensemble approaches, such as classifier chains (CC) and label powersets (LP), have been proposed. CC extends BR by taking some label correlation into account. It works by feeding the predictions of earlier classifiers as features to the latter classifier. However, the CC algorithms suffer from the issue of label ordering, as classifiers with different

chain positions receive different levels of information. LP-based classifiers use subsets of label-sets as class identifiers where each unique set of labels for an MLD is considered for a single label. On datasets with a large number of label combinations, LP has a drawback of ending up with a large number of represented classes and few samples to train on. Random k-label-sets (RAkEL) [303] is an improvement to avoid the problem of the LP method within the large number of unique label-sets. It constructs an ensemble of LP classifiers, and each one is trained using a different small random subset of labels. The classes are then determined by a voting strategy using a threshold. The RAkEL method takes label correlation into account and has lower complexity than the LP method. All three approaches (BR, CC, and LP) are grouped under problem transformation methods, where the MLC problem is transformed into a binary or multi-class problem.

Ranking by pairwise comparison (RPC)[304] creates a pairwise transformation of the multi-label dataset into $L(L-1)/2$ binary label datasets ,one for each pair of labels $(\lambda_i, \lambda_j), 1 \leq i < j \leq L$. On each dataset, a model is trained based on examples annotated by at least one of the labels, but not both. Calibrated label ranking (CLR) [174] extends RPC by introducing one additional virtual label, which indicates the boundary (separation point) between relevant and irrelevant labels. For classifying a new instance, each binary classifier is invoked to vote and predict one of the two labels. Finally, classifiers are evaluated, and the labels are ranked according to their sum of votes. This way, it manages to solve both the MLC and MLR (multi-label ranking) tasks. MLkNN (Multi-label K nearest neighbors) is an adaptation method of the K nearest neighbors (KNN) lazy learning algorithm to multi-label data [175]. MLkNN uses the same basic principle with KNN, except that MLkNN uses a Bayesian approach of prior probability and posterior probability to specify the relevant label-sets, for instance.

### 5.2.4 Imbalance Quantification

The imbalance quantification method designed for single label (binary /multi-class) classification assumes the ratio of minority to majority class as imbalance measure, which is not suitable for multi-label classification. Learning from an imbalanced MLD is a more complex problem in MLC due to the large label space when considering all possible label combinations. Several resampling approaches have been proposed to reduce the imbalanced problem in an MLC (see chapter 3, section 3.4). One of the main challenges of balancing label distribution through resampling methods is that

adding new instances with minority labels also increases the frequency of labels, which are already majority ones. Similarly, removing instances from majority labels will lead to the loss of minority ones [192].

To solve this problem, we proposed a hybrid approach that combines SMOTE (synthetic minority oversampling) with Tomek links named ML-TLSMOTE (Multi-label SMOTE with Tomek links). ML-TLSMOTE can be used as a heuristic-based approach and combination of preprocessing methods whereby the SMOTE and Tomek links (T-link) cleaning method are applied sequentially. SMOTE is applied first to generate synthetic instances of minority labels, and subsequently, T-link, which is used as a post-process cleaning step, is applied to the dataset composed of the original and new synthetic instances with majority labels. Each method, which works well for single-label classification, is adapted to the multi-label scenario to take advantage of their hybrid version. While the T-link method has been successful when used as a secondary preprocessing measure, the T-link cleaning method following SMOTE (i.e., ML-TLSMOTE) doesn't appear to have been explored, particularly in the multi-label scenario. Hence, the aim is to extend and combine each single label resampling method to be used as a solution for multi-label problems. It is plausible that this combination could yield better results in addressing the imbalanced problem in the multi-label learning paradigm by taking their strong points and reduce their weakness. T-link [205] is an enhancement of the nearest neighbor rule [305], which heuristically removes only the noisy or boundary instances of the two classes. The basic idea of the T-link algorithm is as follows:

1. Let i be an instance of class A and j be an instance of class B.
2. Let d(i, j) be the distance between i and j.
3. (i,j) is a T-link, if for any instance m, d(i,j) $<$ d(i,m) or d(i,j) $<$ d(j,m) . If any two examples are T-link, then one of the instances is noise, or both instances are located at the border of the class.
4. Remove noise or border points
5. Repeat steps 1 to 3 until all possible pairs of classes are processed.

For a dataset with two target class values, a T-link is a pair of samples that are (1) nearest neighbors of one another, and (2) have different target class values [205]. Instances that belong to T-link pairs are likely to be either noise points or points that lie close to the optimal decision boundary.

Eliminating these points can result in more well-defined class groups in the training data, which can lead to better classifiers [186]. T-link could be used as an undersampling technique or as a post-process cleaning step [306]. If it is used as an undersampling technique, only the samples from the majority class are removed. If it is used as a post-process cleaning step, samples from both the majority and minority classes are removed. In this study, we used T-link as a post-process cleaning step for two main reasons: (1) to reduce the imbalance between labels by removing instances that are associated with the majority labels, and (2) at the same time to clean up the non-associated instances of labels that were added as a result of the SMOTE procedure, so that the imbalance within a label can be reduced or will not goes to the extreme. In addition, after applying SMOTE on minority labels, the class groups of labels may not be well defined or overlapped due to the invasion of synthetic samples. Therefore, a data cleaning stage is desirable to clean up the borders between each class.

SMOTE [207] technique is an oversampling method where minority class is oversampled to generate new instances using an interpolation technique. The basic idea is to create new samples that are located anywhere on the line that joins together each of the minority class samples and all (or same) of its k nearest neighbors (KNN). The Euclidean distance function is the most commonly used distance metric in KNN. The synthetic samples in SMOTE are generated using the following steps:

1. Choose the feature vector of the current sample (minority class sample)
2. Calculate its k nearest neighbors and randomly select the feature vector of one of these nearest neighbors.
3. The new instances are generated by interpolation technique (e.g., the difference between the selected feature vector of the sample and its selected nearest neighbor)
4. Multiply the result obtained in step 3 with a random value between 0 and 1 and add this vector to the feature vector of the current sample. This causes the selection of a random point along the line segment between two specific feature vectors.
5. The new vector will be the synthetic sample. Repeat these steps until it reaches the number of instances to be generated.

In the multi-label scenario, SMOTE produces a set of instances in which each minority label appears. Each minority instance will be the seed (i.e.,

used as a reference point) for a new synthetic sample. The set of features and label-sets appearing in the reference instances will also be added for the new instances. This hybrid version of SMOTE and T-link (ML-TLSMOTE) is used to reduce the imbalance among the labels. This approach works on the individual minority and majority labels in the MLD. The set of majority and minority labels can be identified based on the imbalance ratio per label (IRLbl), Max imbalance ratio (MaxIR), Coefficient of variation of IRLbl (CVIR), and mean imbalance ratio (MeanIR). These imbalance measures are defined in chapter 3, section 3.3.4.

As it is declared in [192], the joint use of MeanIR and CVIR measures represent if an MLD is imbalanced or not, whereas IRLbl is important to evaluate the imbalance level of each individual label. Any MLD with a MeanIR value higher than 1.5 and a CVIR value of 0.2 should be considered as imbalanced [192]. To come up with balanced labels using ML-TLSMOTE, all instances that are both associated and non-associated with the current minority label are considered for SMOTE; at the same time, these instances should be non-associated with other label combinations. Then, T-link is applied for each majority label to make some adjustments between the classes of each individual label. The algorithm for ML-TLSMOTE is shown in Appendix E. Generally, the joint of use of SMOTE and T-link algorithm is designed to remove the imbalance between the labels and also to reduce the imbalance within the labels. Thus, there are four main aspects to solve the imbalance problem using ML-TLSMOTE in accordance with the following steps:

1. Majority and Minority labels selection: First, the set of majority labels and set of minority labels are identified from the MLD with the help of MeanIR and IRLbl. Labels whose IRLbl is smaller than MeanIR can be considered as majority labels, while labels whose IRLbl is greater than MeanIR are considered as minority labels [192].

2. Multi-label SMOTE: The MLD has more than one minority label. Therefore, each instance associated with each minority label (i.e., instances with 1's for the minority label), but non-associated (value 0's) with other label combinations are oversampled using SMOTE. Selecting active (associated) labels of minority instances and non-associated labels of majority instances help to increase only the frequency of rare labels without cloning the instances that are linked to the majority labels.

3. Multi-label T-Link: in this step, instances that are linked to the majority labels are treated through the T-link cleaning method. T-link allows removing only the noisy or border samples of the majority labels. Removing T-link points can result in more well-defined class clusters in the training data, which can improve the performance of classifiers [186].

4. Reassess the imbalance level: finally, the IRLbl, MeanIR, and CVIR will be recalculated to check if the preprocessed MLD is balanced. At this stage, the MLD could have a more balanced label distribution and would be easier to process by the MLC algorithms.

The algorithm for pre-processing the dataset is implemented using Python. The experimentation part of the study was carried out using the MEKA library [243]. MEKA is an open-source framework for multi-label learning and evaluation, which has been employed for training and comparison of the multi-label classifiers.

## 5.3 Results and Discussions

From the total dataset of 1,095,612 older adults aged 65 years and over, we extracted 105,962 instances to a new MLD, where each instance is associated with at least one active label of the label-set. Several experiments were conducted using the original data, which is basically an imbalanced MLD, and also using the resampled MLDs through SMOTE, T-link, and ML-TLSMOTE for testing the MLC algorithms.

Table 5.3: Characteristics of the MLD before and after applying resampling algorithms

| **Resampling Methods** | **MaxIR** | **MeanIR** | **CVIR** | **Card** | **Dense** |
|---|---|---|---|---|---|
| Without resampling | 5.9 | 2.85 | 0.8 | 1.38 | 0.23 |
| T-link (Under sampling ) | 1.7 | 1.42 | 0.18 | 0.5 | 0.08 |
| SMOTE (Oversampling) | 1.42 | 1.25 | 0.13 | 2.02 | 0.34 |
| ML-TLSMOTE (Hybrid version) | 1.4 | 1.17 | 0.12 | 1.8 | 0.3 |

ML-TLSMOTE is the SMOTE followed by the T-link cleaning method for improving the performance of multi-label algorithms on the dataset with overrepresented and underrepresented labels. Once the resampling approaches were applied to the extracted data, the imbalance level of the preprocessed MLD was re-evaluated. Table 5.3 presents the MaxIR, MeanIR,

and CVIR values for each dataset along with the distribution of the labels.When we compare the imbalance measures (MaxIR, MeanIR, and CVIR) of resampled data with the one without resampling (first row) in Table 5.3, it can be seen that a general improvement in the imbalance levels has been achieved.

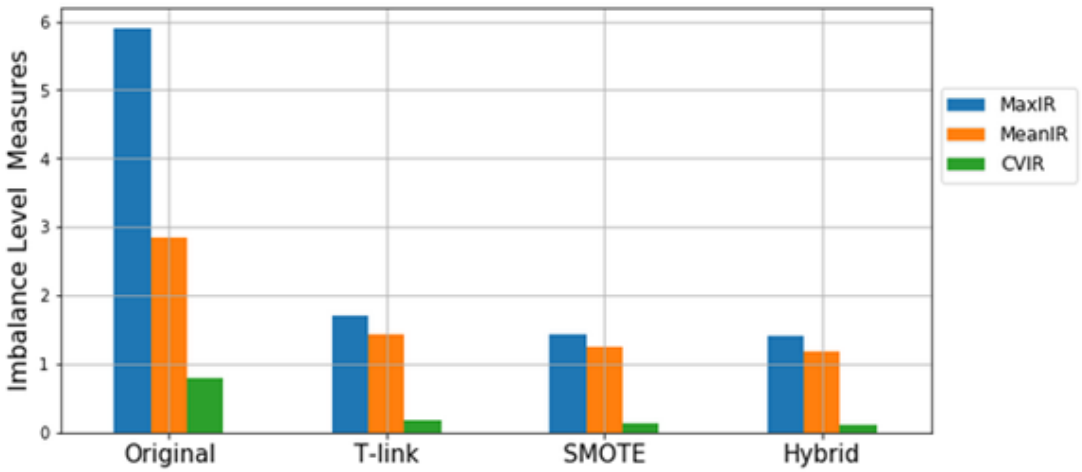


Figure 5.1: The behavior of the data after applying the resampling approaches

The average imbalance level of the data after applying ML-TLSMOTE is MeanIR=1.17 and CVIR=0.12, which gives evidence that the imbalanced problem has been much reduced in the data. Although it might not always be the case, MaxIR, MeanIR, and CVIR are lower after applying resampling approaches with the lowest value (i.e., the better result) found in the hybrid of oversampling and cleaning methods (ML-TLSMOTE). This implies that the ratio between the most frequent and the least frequent labels have been reduced in the data. Figure 5.1 shows the behavior of the original dataset after applying the resampling approaches. To understand how these imbalance levels could influence the classification results, we used various multi-label classifiers. In this experiment, six different MLC algorithms were chosen: Binary Relevance (BR) [307], Classifier Chains (CC) [101], Label Powerset (LP), Random k-label sets (RAkEL) [303], Calibrated Label Ranking (CLR) [174] and MLkNN [175]. All the experiments were conducted using 10-fold cross-validation. Different base classifiers such as decision trees, random forest, and naïve Bayes were used for each multi-label algorithm.

Although the change in imbalance level will not necessarily imply better

classification results, it has been observed that the lower the values of the imbalance levels, the better the performance of the MLC algorithms. In this regard, the performance of MLC algorithms has shown improvement when using SMOTE and ML-TLSMOTE with respect to the different multi-label performance measures.

The evaluation of models in MLC differs from the traditional single-label classification [175]. It requires a special approach in order to consider performance over all labels. In this study, the average precision, Hamming loss, ranking loss, and Hamming score were used to evaluate the performance of different MLC models. Figure 5.2 shows the comparison of oversampling (SMOTE), under-sampling (T-link), and hybrid approach (ML-TLSMOTE) on various MLC algorithms using the Hamming loss, ranking loss, and average precision. Hamming loss discriminates the capability of the algorithm to identify the presence of frailty in terms of adverse health outcomes. Ranking loss discriminates how well the algorithm ranks labels, which allow understanding the type of patient outcomes that have a strong expression, giving an indication of where to act promptly. Average precision allows knowing the percentage of correct positive predictions.

Although the MaxIR, MeanIR, and CVIR are reduced after applying T-link on the original data, the classifiers built using the dataset created with the T-link under-sampling technique showed a very poor result. One of the reasons for this could be the fact that T-link can be more robust when used as a secondary pre-processing method following SMOTE than when it is used alone as an independent technique. The other possible and obvious reason could be the type of data used in the study and the distribution of the classes or labels in the dataset. As shown in Figure 5.2, each classifier was applied to the data that has been pre-processed through SMOTE and T-link to compare the results with the proposed ML-TLSMOTE method. It is plausible that more balanced MLDs could yield higher classification performance than the less balanced ones. ML-TLSMOTE has the lowest score of Hamming loss and ranking loss in all classifiers followed by SMOTE, indicating the classification results obtained with ML-TLSMOTE are better than the results found through SMOTE and T-link individually. SMOTE has the best result next to ML-TLSMOTE, while T-link showed lower results, even worse than the results of the original imbalanced MLD for some classifiers.

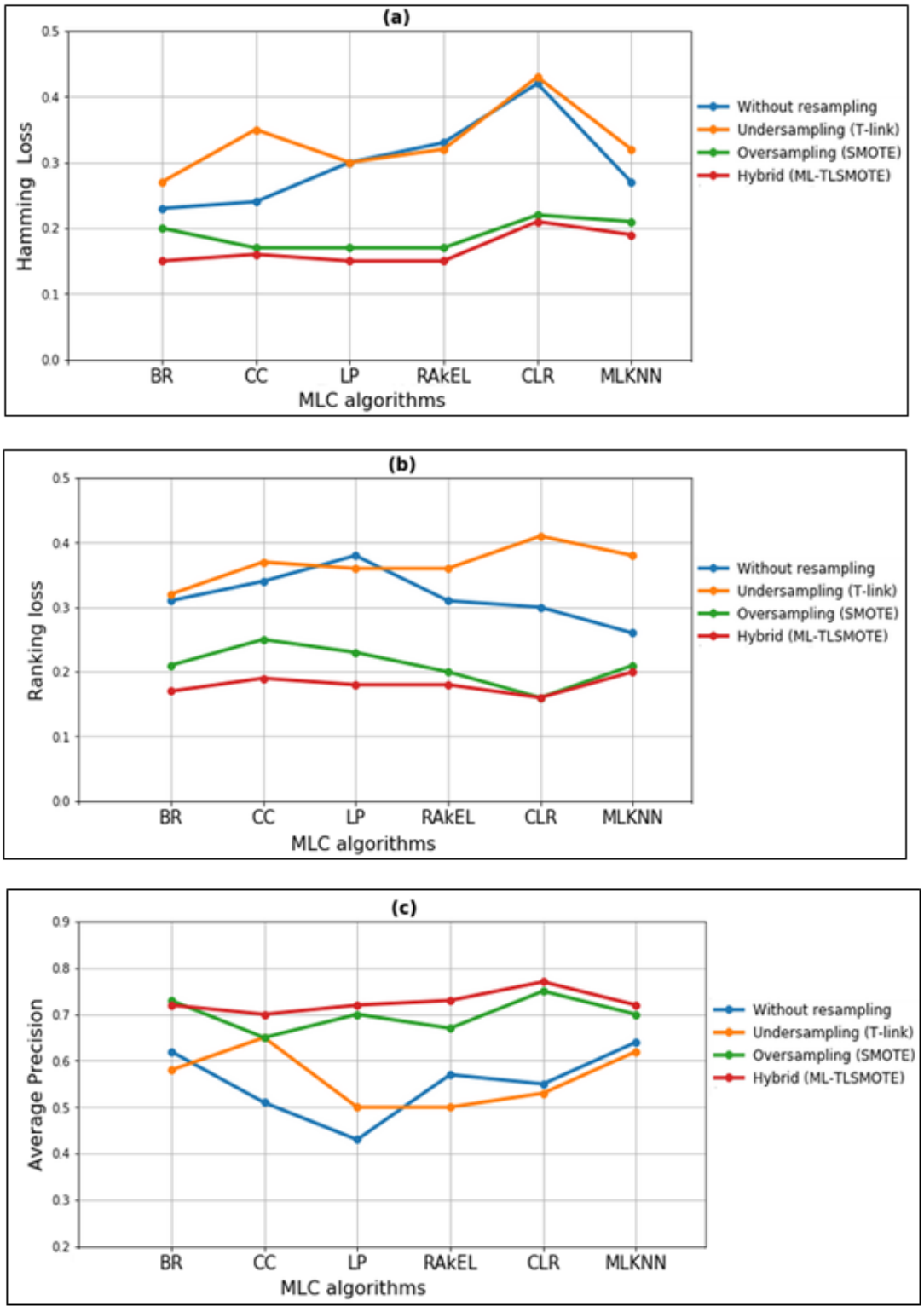


Figure 5.2: Comparison of resampling methods through various classifiers using Hamming loss (a), ranking loss (b), and average precision (c).

Tables 5.4 and 5.5 present the detailed results of 10 fold cross-validation produced by six different MLC classifiers. Table 5.4 shows the results of classifiers on the original imbalanced MLD, and Table 5.5 presents the results after applying ML-TLSMOTE. The results show that the ML-TLSMOTE significantly improves the performances of classifiers in all evaluation metrics.

Table 5.4: Results before applying ML-TLSMOTE (original MLD)

| **Metrics** | **BR** | **CC** | **LP** | **RAkEL** | **CLR** | **MLkNN** |
|---|---|---|---|---|---|---|
| Hamming Score | 0.77 | 0.76 | 0.7 | 0.67 | 0.58 | 0.73 |
| Average precision | 0.62 | 0.51 | 0.43 | 0.57 | 0.55 | 0.64 |
| Hamming loss | 0.23 | 0.24 | 0.3 | 0.33 | 0.42 | 0.27 |
| Ranking loss | 0.31 | 0.34 | 0.38 | 0.31 | 0.3 | 0.26 |

Table 5.5: Results after applying ML-TLSMOTE

| **Metrics** | **BR** | **CC** | **LP** | **RAkEL** | **CLR** | **MLkNN** |
|---|---|---|---|---|---|---|
| Hamming Score | **0.85** | 0.84 | **0.85** | **0.85** | 0.79 | 0.81 |
| Average precision | 0.72 | 0.7 | 0.72 | 0.73 | **0.77** | 0.72 |
| Hamming loss | **0.15** | 0.16 | **0.15** | **0.15** | 0.21 | 0.19 |
| Ranking loss | 0.17 | 0.19 | 0.18 | 0.18 | **0.16** | 0.2 |

To further analyze the results shown in Tables 5.4 and 5.5, and to see if there is any significant difference between the classifier results in terms of all evaluation metrics, a Wilcoxon signed-rank test is performed on the results obtained before ML-SMOTE versus after ML-SMOTE. The test is applied to each performance metric across the six algorithms with a significance level of 0.05. The obtained P-value is less than 0.05, which shows that applying ML-TLSMOTE significantly improves the classification results, and the difference of results before and after ML-TLSMOTE is big enough to be statistically significant.

We also noticed that the use of different base classifiers for each MLC algorithms had shown a more significant effect in the variation of classification results. Classifier chains have shown the best results in terms of Hamming loss and average precision when using the random forest as a base classifier, while RAkEL has achieved the best results using SMO (Sequential Minimal Optimization) as the base classifier. MLkNN used naïve Bayes as the base

classifier. Among the MLC algorithms, RAkEL, LP, and BR achieved the best performance in terms of Hamming loss with ML-TLSMOTE (Table 5.5). They also have a similar value of Hamming score (85%) with an average precision of 73% for RAkEL and 72% for BR and LP. The Hamming score and Hamming loss capture the fraction of labels that are correctly and incorrectly predicted, respectively.

Ranking loss measures the average fraction of labels that are ordered incorrectly. For example, the ranking loss of RAkEL in Table 5.5 is 0.18, which means that 18% of the label pairs are wrongly ordered for instances. With ranking evaluation measures, CLR outperforms the other algorithms which rank the relevant labels higher than irrelevant labels efficiently based on the pairwise comparison of labels. CLR has also achieved the best result in average precision (77%), while the multi-label variant of KNN (i.e., MLKNN) showed poor performances in Hamming loss and ranking loss as compared to BR, CC, and LP. The BR doesn't consider label correlation. It uses six classifiers separately, which is equal to the number of labels in the frailty dataset. Overall, from the analysis of results, it can be concluded that the ML-TLSMOTE has a more positive influence over classifiers, such as BR, CLR, CC, LP, RAkEL, and MLKNN.

Finally, with efficiency measures, the computational complexity of BR, CLR, and LP depends on the complexity of the base classifier and the parameters of the learning problem [308]. We observed that using tree-based methods as a base classifier (e.g., C4.5) is more efficient than the SVM-based methods. The BR algorithm, which builds separate models for each label, is the simplest one. CLR is the next least complex algorithm, requiring |L| number of BR models and additionally |L|*(|N|-1)/2 one against one model. LP is relatively the most sophisticated algorithm, since it trains a multi-class classifier, with the number of classes being equal to the number of distinct label-sets in the MLD. The computational complexity of MLKNN is |L| times the computational cost of computing K nearest neighbors. Training MLKNN model is linear with the size of the training dataset and the length of the data vector.

## 5.4 Conclusions

Detecting frailty in elderly people represents an essential research problem, and there is potential to prevent frailty and intervene early. In this study, MLC was developed for the purpose of predicting multiple outcomes of

frailty conditions: mortality, fracture, disability, medical emergency admission at the emergency department, urgent hospitalization, and preventable hospitalization. MLC models are valuable tools to construct a predictive model that considers the prediction of multiple outcomes and interventions in an unseen patient. The study consists of two major parts: the first is addressing the imbalance problem in an MLC. ML-TLSMOTE was proposed to reduce the imbalance between labels and to improve the performance of MLC algorithms. The results of the experiment show that ML-TLSMOTE was an efficient approach as compared to SMOTE or Tomek links. The second part presents a comparative study of six MLC algorithms (BR, ECC, LP, CLR, RAkEL, and MLKNN). RAkEL achieved the best performance in terms of Hamming loss, while CLR showed the best value of the ranking loss.

As future work, there are three problems that need further investigation in our study. The first is the dimensionality reduction (feature selection or extraction) to optimize and improve the performance of the training models, which is one of the challenging topics in the MLC task. Second, in the advances of wearable and sensor technologies, many elderly people with frailty can use intelligent wearable sensor equipment to monitor the physiological signals; thus, it is essential to collect and analyze real-time data from wearable sensors to make a more accurate risk assessment. Finally, with the development of personalized healthcare, there is a need to study and build a personalized frailty risk prediction model for elderly care.

# Chapter 6

# Cluster Analysis and Its Validation: Towards Improving Health Conditions of Elderly

This chapter presents two main sections on clustering problem. Section 6.1 presents a cross-validation approach to validate clustering results. It proposes a new cluster validity index based on information from multiple labels to measure the effectiveness of the clustering algorithm through exploiting and adjusting the root mean squared error (RMSE). The study validates the proposal through the k-fold cross-validation analysis of some challenging multi-label datasets. Section 6.2 presents cluster analysis aiming to identify homogenous groups of elderly patients aged 65+ years and above. This section examines the various dimensionality reduction and clustering methods to achieve optimal clusters. The resulting clusters were evaluated using various validation techniques, including the validation criteria proposed in section 6.1.

## 6.1 Cross-Validation Approach to Evaluate Clustering Results using Multi-label Datasets

### 6.1.1 Introduction

This subsection first introduces an overview of clustering problem, validation techniques, and multi-label data and the existing challenges of the clustering task. Then the proposed approach to cluster validation is described in detail in section 6.1.2. Finally, the experiments and results are

presented in the subsequent sections.

## Overview of Cluster Analysis

Unsupervised learning aims to find the underlying structure or the distribution of data. It is an important area in the domain of machine learning, where the labels for the data examples are not necessarily required for model building. The main tasks in unsupervised learning include cluster analysis [77, 309], building self-organizing maps (SOM)[79], representation learning [80], and density estimation [81]. Cluster analysis, the main focus of this study, is a central task for grouping heterogeneous data points into a number of more homogenous subgroups based on distance, or naturally occurring trends, patterns, and relationships in the data. The formation of homogenous or heterogeneous grouping (or clustering) structure from a complex dataset requires a measure of 'closeness' or 'similarity.' In clustering, the definition of similarity is highly dependent on the applied distance function between the data objects. The choice of similarity measure can be considered based on the type of the variable used to cluster objects (continuous, discrete, binary), the type of measurements (nominal, ordinal, ratio, interval), and subject matter knowledge. The most commonly used distance measure in most clustering algorithms is the Euclidian distance [310]. Other measures include Minkowski's distance [311],Cosine distance[312],S-distance [313] ,etc.

The clustering problem has a clear goal of finding distinct groups or 'clusters' within the dataset. However, the notion of a 'cluster' has not been precisely defined, which has driven to the development of several clustering paradigms and several clustering algorithms within each paradigm[314]. The existence of different types of clustering algorithms poses difficulties in selecting the best algorithm for a particular task. Independent of the type of algorithm used, Kleinberg [116] proposes three properties that an ideal clustering algorithm should have so that it can be considered good: scale invariance, consistency, and richness. Scale invariance indicates that the clustering algorithm does not change its results when all distances between points are scaled by a constant factor. A clustering process is considered to be consistent when the clustering results do not change if the distances within clusters decreases and/or the distance between clusters increase. The richness criteria mean that the clustering function must be flexible enough to potentially produce any arbitrary partitions of the input dataset. According to Kleinberg's impossibility theorem [116], no clustering algorithm satisfies

all three requirements simultaneously. This implies that it has been very difficult to develop a unified framework for validation of clustering methods and to reason about it at a technical level.

### Multi-label Data

Several types of research in machine learning deal with the analysis of single-label data, where training instances are associated with a single label $\lambda$ from a set of disjoint labels $L$. However, training samples in several application domains are often associated with a set of labels $Y \subseteq L$. Such datasets are called multi-label data. Multi-label datasets have been popular in various domains, such as protein function classification, medical diagnosis, emotion recognition, text classification, etc. For instance, a medical patient may be affected by more than one chronic disease: diabetes, hypertension, and fatty liver. We can cluster the patients into distinct groups, each with specific characteristics, and then the burden of these unwanted outcomes (diabetes, hypertension, fatty liver, etc.) can be identified to provide tailored interventions in each cluster. One of the common trends for solving supervised learning through the use of multi-label data is decomposing the multi-label problem into binary classification problems [315]. In unsupervised learning, we can use the labels information of the multi-label data for evaluation of the clustering algorithm. In this study, we used features for forming clusters and class labels for performance evaluation.

### Cluster Validation

Cluster validation is one of the most important and challenging parts of cluster analysis, which involves the objective and quantitative assessment of clustering results [309]. One of the problems in cluster validation is that there is no clear notion as to what exactly the 'prediction error' is. Because of that, clusters are sometimes validated by ad hoc methods based on the application area. Due to the absence of the ground truth and the nature of the problem, cluster validation has not been well developed [316]. As a result, evaluating the performance of a clustering algorithm is not an easy task. Commonly, the evaluation process depends on the algorithm used to obtain clustering results, which resulted in the development of multiple evaluation techniques. Various methods have been suggested in the literature for cluster validation, including external validation, internal validation, relative criteria, and stability based approaches.

**External Clustering Validity Methods**: The external validation index

uses prior knowledge, such as externally provided class labels, to evaluate results of cluster analysis. External clustering validity approaches, such as Rand Index [317] and normalized mutual information [318], is used to measure the quality of clustering results by comparing the generated cluster labels with the pre-existing clustering (reference labels) structure, i.e., ground truth solution. If the result is in some way similar to the reference, the final output is regarded as a "good" clustering. The external validation is straightforward when the closeness between two clusterings is well-defined. However, it has a basic caveat that the reference result is not given in most real-world applications. Therefore, external evaluation is generally used for synthetic data and for tuning clustering algorithms [319].

**Internal Cluster Validity Methods**: these are used to assess the goodness of the clustering structure without reference to the external information, using only the data themselves. Internal clustering validity methods measure the quality of clustering based solely on information intrinsic to the data; as a result, they have great practical application and numerous criteria have been proposed in the literature, such as Silhouette analyses [320], Calinski–Harabasz index [321], Davies–Bouldin [322]. The internal criteria are the most commonly used evaluation methods designed to compute the ratio of within-cluster scattering (compactness) and to between-cluster separation. Measures that grouped under this category have been designed for the validation of convex-shaped clusters ( such as globular clusters), and fail when applied to validate non-convex clusters [323].

**The relative approach**: is performed by comparing two sets of clusters (usually built with similar algorithms but with different parameter settings) to determine which one is better. It's generally used for determining the optimal number of clusters.

**Clustering Stability Approach**: clustering stability measure is a slightly different approach used to assess the similarity of clustering solutions obtained by applying the same clustering algorithm on multiple independent and identically distributed samples. The intuitive idea behind the stability approach is that if we repeatedly sample data points from the population and apply the candidate clustering algorithm, then a good algorithm should produce clusterings that do not vary much from one sample to another [324]. In other words, the algorithm is stable with respect to input randomization. There are several studies to validate clusters by stability criteria [325, 326]. In general, the existing validation criteria are useful for such tasks as determining the correct number of clusters in the dataset, verifying whether

the clusters obtained are meaningful or are just an artifact produced by the algorithms, justifying why we choose some algorithms instead of others or assessing the quality of clustering solutions. However, in the literature, there is still a lack of methods to measure the ability of the clustering algorithm to predict cluster memberships for new data points.

Generally, evaluating clustering results has been historically expressed as the most challenging topic [327]. In fact, Jain and Dubes [328], in their classic book on clustering, stated that:

*"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage".*

Despite achievements observed in this particular area over the last several decades, it is highlighted that the above statement still remains true. This has motivated us and many other researchers in the area to study, develop, and propose methods to the validation of clustering results.

### Objectives and Contributions

The primary aim of the study under this section is to measure the performance of a clustering model to predict cluster labels for new data points, given that the model is already constructed from the training data. For example, we have three existing clusters, C1, C2, and C3, and a new data point D1. The clustering model should assign D1 to one of the clusters, say C2. In this case, we want to know 'how good is the model on new data?' i.e., to what extent the model has correctly assigned D1 into C2.

The cluster validation idea presented in this study is different from the existing methods in that it focuses on measuring the prediction strength of a clustering algorithm by using the cross-validation procedure. The k-fold cross-validation method is used for simulating the situation when we have built the clustering model on some previously available data, and then we want to assign new data points to the previously built clusters. The prediction strength concept presented here, similarly, as the stability of the clusters, can be used for assessing the performance of a clustering method. Clustering stability results are mostly obtained based on perturbations introduced to the input data, such as sub-sampling or the addition of noise.

Unlike in the other studies, the prediction strength of an algorithm introduced here is measured by incorporating information from several labels of multi-label data. Namely, the probability of occurrence of the labels in the training and testing data is calculated for each cluster. If label probabilities in the training and testing data are similar, the clustering can be considered as a good one. Thus, this study assumes that the clusters are already formed from the training data, and the aim is to measure how well the clustering model predicts the corresponding cluster labels for the test data based on their membership on the clustering results obtained from the training data.

This approach is motivated by medical applications in which we would like to assess the probability of various health problems in different patient groups. For example, the labels for the chronic dataset are diabetes, hypertension, and fatty liver. Once the clusters are formed, the probabilities of the occurrence of these labels, i.e., diabetes, hypertension, and fatty liver are estimated in each cluster and compared between the training set and the test set. The aim is to measure how well we can predict the probabilities of these three outcomes in new patients (i.e., in the test data) based on their membership in the training clusters. In this study, the k-fold cross-validation procedure is used to simulate such a scenario.

The k-fold Cross-Validation (CV) is one of the most commonly used model evaluation procedures in supervised learning. Unfortunately, it is challenging to apply CV to unsupervised learning, for example, to clustering validation. Here, the k-fold CV procedure is adapted, by using labels from a multi-label dataset, to be applicable to unsupervised learning (i.e., clustering) for evaluating the performance of clustering algorithms. Following the k-fold cross-validation approach, the input data is randomly divided into k parts, of which k-1 parts are used to construct the model, and the remaining part is used as an evaluation set. Then, the prediction strength is used as a statistic for clustering stability. Thus, here we propose the use of the k-fold cross-validation procedure for evaluating the prediction strength of the clustering model using the information acquired from multiple labels.

The contributions of this study are: (1) a new cluster validity index is proposed that uses the information from multiple labels to evaluate the quality of clustering algorithms; (2) the study validates the proposal through the cross-validation analysis of some challenging multi-label datasets; (3) the root mean squared error (RMSE), which is the most frequently used measure of the differences between values in regression problem, is exploited

and adjusted to be used as a cluster validity index; (4) this study shows that the proposed method can be used to measure the ability of a clustering algorithm to predict the cluster membership for new data.

### 6.1.2 Proposed Method

Given a particular clustering result, one can predict cluster membership for new data based on a clustering model built on training data. This is not always easy for all types of clustering algorithms. For example, it is hard for density-based clustering algorithms (e.g., DBSCAN) to predict a cluster for the new data points, because the new data points may change the underlying clustering structure. For centroid-based cluster algorithms (e.g., k-means clustering), however, the prediction of a cluster for new data points is relatively easy since it only requires finding the minimum distance of a new data from all cluster centers and then updating the cluster center of that cluster. Hence, k-means clustering is employed to test the proposed method in this study. Recently, several techniques have been proposed to improve the standard k-means algorithm for high dimensional datasets, such as the Entropy Regularized Power k-Means [329], sparse k-means [330] and others [331]. The proposed k-Fold CV for unsupervised learning can also be applied to these modified versions of the k-means algorithms.

Assigning new data points to existing clusters that are constructed through the training data is considered to be an important practical application. However, very little practical guidance is available to measure the prediction strength of the constructed model to predict the cluster membership of a new data point. Prediction strength is a global measure forcing all clusters to be stable, as it uses the minimum value of cluster similarity over all clusters [127]. This thesis proposes a k-fold cross-validation procedure followed by the root mean squared error (RMSE) or the mean absolute percentage error (MAPE) to evaluate the prediction strength of the clustering algorithm. RMSE and MAPE are the most commonly used error measurements in statistics. In prediction tasks, RMSE indicates the absolute fit of the model to the data, i.e., it is used to compare how close the observed data points are to the predicted values of the model. MAPE is the average magnitude of the difference between predicted and actual values in percentages, without considering their direction, that is, since absolute percentage errors are used, the positive and negative errors are not canceling each other. In clustering validation, these two metrics can be used to measure the average distance between the data points and their cluster centers [332, 333]. The

smaller the RMSE/MAPE, the better the prediction results.

At each iteration of the k-fold CV procedure, one fold is used as the test set and the remaining folds as the training set. The training set is presented to a clustering method, giving a partition as a result (training partition). Then, new data points are assigned to the clusters in the training partition based on the minimum distance from all the cluster centers. The CV method allows calculating the quality measure expressing the difference between the probability of occurrence of the outcomes (i.e., labels) in the training data and in the test data assigned to the same cluster. Once the clusters are formed using the training part of the data, the probability of occurrence of the labels in the training set, and in the testing set in each cluster will be assessed and analyzed. This is similar to estimating the probability that an outcome will occur, given that a sample belongs to a certain cluster, mathematically written as P(outcome|cluster). For instance, in the chronic disease dataset, one can estimate a probability of the risk of having hypertension in each of the generated clusters. Below, we describe the k-fold cross-validation procedure used to calculate a quality measure for a clustering model.

Let: $L = \{\lambda_1, \lambda_2, ..., \lambda_q\}$ : the set of all labels in a multi-label dataset
$q = |L|$: the number of labels in the multi-label dataset.
$k$: the number of folds in the cross-validation procedure,
$C$: the number of clusters generated by the clustering algorithm.

Because we calculate label probabilities separately for each cluster $i$ in each of the cross-validation folds $j$ we denote these probabilities without using the number of the cluster nor the number of the fold in order not to clutter the equations:
$y_m, m = 1, \ldots, q$: the probability that a sample from the training dataset assigned to cluster $i$ has the $m^{th}$ label
$\hat{y_m}, m = 1, \ldots, q$: the probability that a sample from the testing dataset assigned to cluster $i$ has the $m^{th}$ label

1. Shuffle the original dataset randomly

2. Split the original dataset into k parts (folds)

3. For each fold $j = 1, \ldots, k$.

(a) Take fold $j$ as the test dataset (each fold, in turn, is used as the test dataset).

(b) Take the remaining folds together as the training dataset.

(c) apply dimensionality reduction (if needed )

(d) apply normalization to dataset (if needed)

(e) Generate clusters on the training dataset.

(f) Assign data points from the test dataset (selected in step 'a') into the corresponding clusters obtained in step 'e'.

(g) For each cluster $i = 1, \ldots, C$ :

i. Compute the probabilities $y_m, m = 1, \ldots, q$ of the occurrence of the labels in cluster i based on the samples in the training dataset, found in step 'e'.

ii. Compute the probabilities $\hat{y_m}, m = 1, \ldots, q$ of the occurrence of the labels in cluster $i$ using the assignment of the points from the test dataset to the clusters, which was obtained in step 'f'.

iii. Compute the root mean squared error ($RMSE_{ij}$) between the probabilities calculated in steps i and ii. Note down the scores/errors as a quality measure for cluster $i$ obtained in fold $j$.

4. When the loop in step 3 finishes (and so every fold served as the test set), take the average over the $k$ folds of the recorded scores for each cluster and/or overall the clusters).

In the context of this clustering validity criteria, RMSE and MAPE are proposed to measure the prediction strength of clustering techniques. RMSE represents the standard deviation of the difference between the probabilities of occurrence of the labels of the training data and the probabilities of occurrence of the labels of the test data in clusters. Intuitively, the RMSE in this study can be understood as the Euclidean distance between the vector of the observed probability scores of labels in the training data and the estimated probability scores of the labels in the test data for a given cluster, averaged by the total number of labels in the data (Eq. 6.1). Similarly, MAPE measures the size of the error between the probability scores of the training set and the probability scores of the test set in percentage terms (Eq. 6.2). RMSE and MAPE are evaluation methods that can be used together to diagnosis the variation in the errors of a clustering algorithm. For cluster $i$ and cross-validation fold $j$, these two measures are calculated as follows:

$$RMSE_{ij} = \sqrt{\frac{\sum_{m=1}^{q} (\hat{y_m} - y_m)^2}{q}} \tag{6.1}$$

$$MAPE_{ij} = (\frac{1}{q} \sum \frac{(y_m - \hat{y_m})^2}{y_m}) \times 100 \tag{6.2}$$

The resulting score obtained through RMSE with k-fold cross-validation across all clusters based on the probability score information from multiple labels, named CVIM in short, can be used as a cluster validity index (i.e., stability index). The better the values of the cluster validity index, the more stable the outputs of the clustering algorithm. High cluster stability is achieved when memberships of the clusters are not affected by small changes in the data set. The RMSE of the clustering algorithm obtained using the k-fold cross-validation can be computed as shown in equation 6.3: let $RMSE_{ij}$ be the RMSE for the $i^{th}$ cluster obtained in the $j^{th}$ fold. The average RMSE for the $i^{th}$ clusters obtained in $k$ fold with $C$ clusters in each fold, denoted by $ARMSE_i$ , can be computed as:

$$ARMSE_1 = (RMSE_{11} + RMSE_{12} + RMSE_{13} + ... + RMSE_{1k})/k$$

$$ARMSE_2 = (RMSE_{21} + RMSE_{22} + RMSE_{23} + ... + RMSE_{2k})/k$$

$$\vdots$$

$$ARMSE_C = (RMSE_{C1} + RMSE_{C2} + RMSE_{C3} + ... + RMSE_{Ck})/k$$

$$Overall\ ARMSE = (ARMSE_1 + ARMSE_2 + ... + ARMSE_C)/C$$

$$Cluster\ Validity\ Index(CVIM) = \frac{1}{C} \sum_{i=1}^{C} ARMSE_i \tag{6.3}$$

Finally, the RMSE based cluster validity index across all clusters is found using equation 6.3. The MAPE based CVIM is also computed in a similar

fashion as the RMSE. The architecture of the proposed method for calculating RMSE and MAPE for each cluster in ten folds of cross-validation is presented in Figure 6.1 for an algorithm generating C = 3 clusters. In the final stage, the average RMSE/MAPE of 10 similar clusters is taken from each fold of cross-validation.

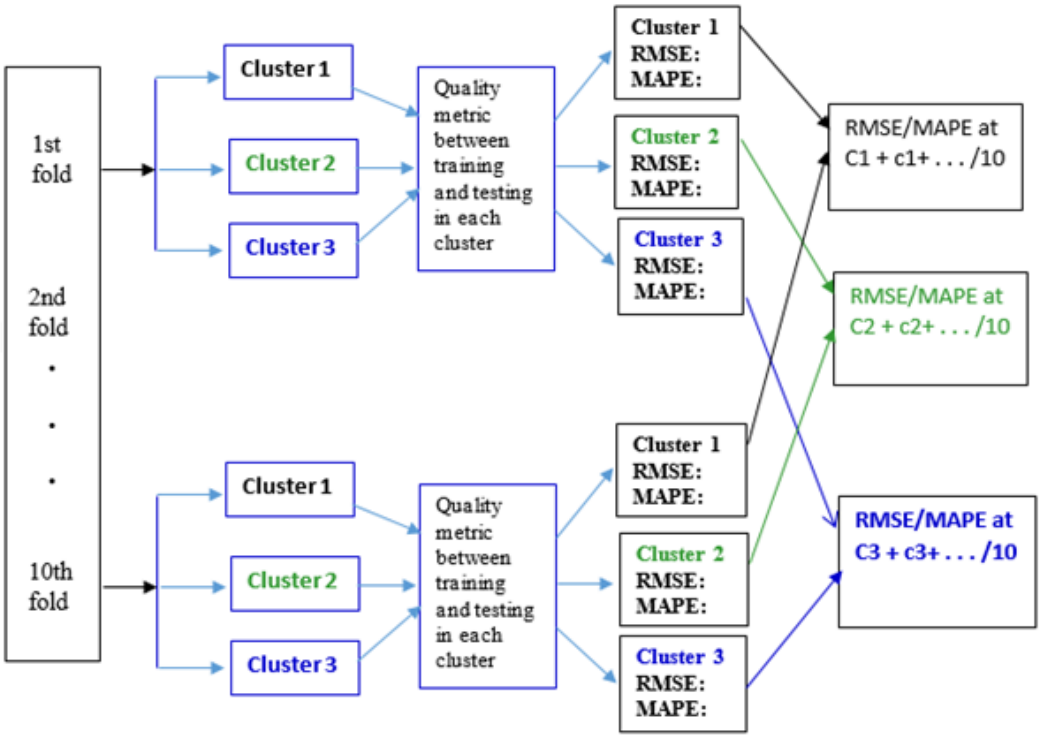


Figure 6.1: The architecture of the proposed method to evaluate a clustering model through 10 fold cross-validation with three clusters at each fold.

### 6.1.3 Experiments

In this section, three public multi-label datasets were used to test the proposed method: the chronic diseases dataset [334], emotions [335], and Yeast [109] datasets. The chronic diseases dataset contains a collection of physical examination records for 110,300 patients with 62 features and 3 class labels. All the input features were used for forming clusters. The class labels (non-clustering variables), which include hypertension, diabetes, and fatty liver, were not used for defining clusters but only for cluster validation. Each record in the data may be associated with more than one of

the class labels. As a result, the probability of occurrence of hypertension, diabetes, or fatty liver in patients of the test data can be estimated in the corresponding clusters. The chronic disease dataset is available online at http://pinfish.cs.usm.edu/dnn/. The Yeast dataset is formed by micro-array expression data and phylogenetic profiles with 2,417 genes. The dataset consists of 103 features with 14 labels, and each gene is associated with a set of functional labels. The emotions dataset contains examples of songs according to people's emotions. The emotions and Yeast datasets were taken from the Mulan Library at http://mulan.sourceforge.net/datasets-mlc.html.

Multi-label datasets, and current data, in general, tend to be more complex than conventional data and need dimensionality reduction. All three multi-label datasets used in this experiment have a large number of features and labels/outcomes. Taking this problem into account, we applied the dimensionality reduction process to convert the dataset into two-dimensional space. The purpose of reducing data into lower-dimensional representation is to visualize and interpret the samples so that such visualization can be used to obtain insights from the data, e.g., to detect clusters and identify outliers. Moreover, a clustering process requires data reduction to obtain an efficient processing time while clustering and avoid the curse of dimensionality. For example, the k-means clustering algorithm often doesn't work well for high dimensional data [336]. There are different techniques proposed in the literature for high dimensional features in clustering [337, 338]. In this study, principal component analysis (PCA)[339], one of the most commonly used technique, was applied as a data dimensionality reduction to convert each dataset into a two-dimensional representation. Emotions and Yeast datasets have large variations within the range of feature values, which can affect the quality of computed clusters. Therefore, after PCA, we applied the normalization technique [340] for Emotions and Yeast datasets to ensure that good quality clusters are generated. Then, k-means clustering [341] was applied to the reduced dataset. All the experiments have been implemented using Python programming language.

### 6.1.4 Results and Discussions

With the help of the Calinski-Harabasz index, three clusters for emotions dataset, four clusters for chronic disease dataset, and five clusters for yeast dataset were identified using the k-means clustering algorithm. A two-dimensional (2D) representation of clustering results for each dataset is

shown in Figure 6.2. Colors of the points represent cluster memberships of the samples.

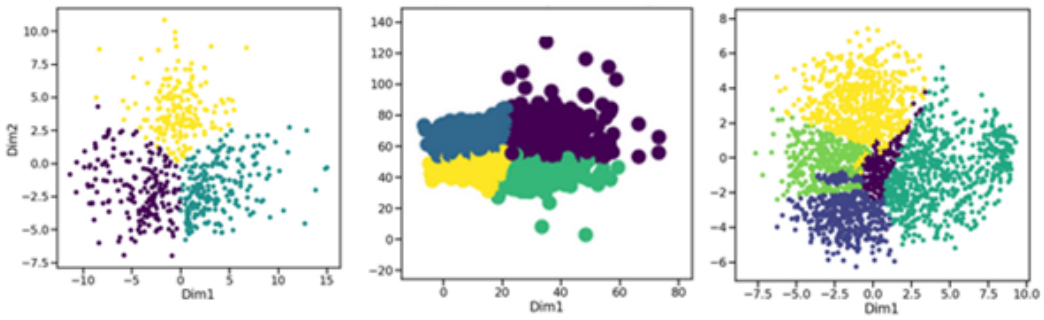


Figure 6.2: 2D visualization of clustering results on Emotions (a), chronic disease (b), and Yeast (c) datasets.

Min-Max normalization method has been applied to Emotions and Yeast datasets to eliminate the large variations within the range of features before the clustering process. For each dataset, the probabilities of the occurrence of each target variable in each cluster have been calculated both in the training and testing part of the data during the cross-validation procedure. We first evaluated the quality of the clusters using the existing internal validity criteria. Silhouette analysis is one of the most popular and effective internal measures which allows evaluating the appropriateness of the assignment of a data object to a cluster by measuring both intra-cluster cohesion and inter-cluster separation. Clusters within the range of 51 to 70% and 71 to 100% respectively indicate that a reasonable and a strong intra-cluster cohesion and inter-cluster separation are found [342]. The silhouette score can take values in the interval [-1, 1]. Negative silhouette values represent wrong data placements, while positive silhouette values better data assignments. Therefore, we want the scores to be as big as possible and close to 1 to have good clusters. In our experiments, the silhouette score has shown good results. The silhouette score for clusters found on emotion, chronic disease, and Yeast datasets were 0.76, 0.82, and 0.69, respectively, indicating that the obtained clusterings were good ones.

As the main objective of this study is to evaluate the prediction performance of the clustering algorithm through a 10-fold cross-validation procedure, the result of prediction performance in terms of RMSE and MAPE are presented for each cluster and across all clusters (i.e., the CVIM value), as shown in Table 6.1. The results represent the strength of the clustering

algorithm to predict cluster labels for the test data. The obtained RMSE and MAPE scores of the clustering results in each cluster of each dataset represent the prediction errors. Figures 6.3 and 6.4 show the RMSE and MAPE of the k-means clustering algorithm applied to each dataset, respectively. The smallest RMSE (i.e, the better result) is found in the Emotions dataset in each cluster, while the highest RMSE was found in the Yeast dataset. This also holds true for the total RMSE across all the clusters (i.e., the CVIM score) on each dataset.

Table 6.1: Performance of a clustering algorithm in each and across the clusters using CVIM

| Dataset | Metrics | Cluster1 | Cluster2 | Cluster3 | Cluster4 | Cluster5 | CVIM |
|---|---|---|---|---|---|---|---|
| Emotions | RMSE | 0.021 | 0.019 | 0.017 | - | - | 0.019 |
| | MAPE | 7.88% | 18.27% | 8.99% | - | - | 11.71% |
| Chronic | RMSE | 0.0361 | 0.0543 | 0.0228 | 0.0282 | - | 0.0354 |
| | MAPE | 5.62% | 5.92% | 7.91% | 12.29% | - | 7.94% |
| Yeast | RMSE | 0.071 | 0.061 | 0.066 | 0.086 | 0.076 | 0.072 |
| | MAPE | 7.49% | 9.36% | 11.59% | 17.34% | 15.34% | 12.22% |

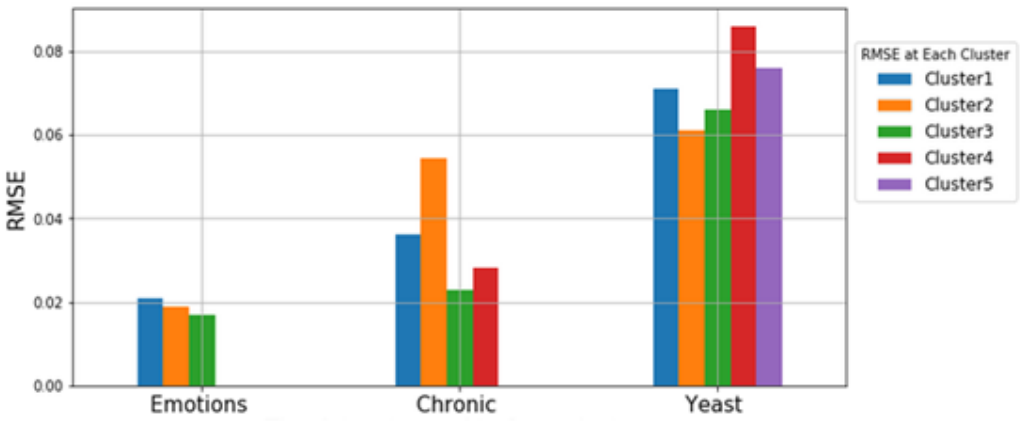


Figure 6.3: RMSE of the clustering algorithm on each cluster in each dataset

Generally, an RMSE close to zero is indicative of the high similarity between the training and testing probabilities. Similarly, low MAPE values indicate good predictions of the occurrence of labels in each cluster across all datasets. The smaller the MAPE, the better the forecast, and more specifically, Lewis's [343] interpretation of MAPE is that a value of less

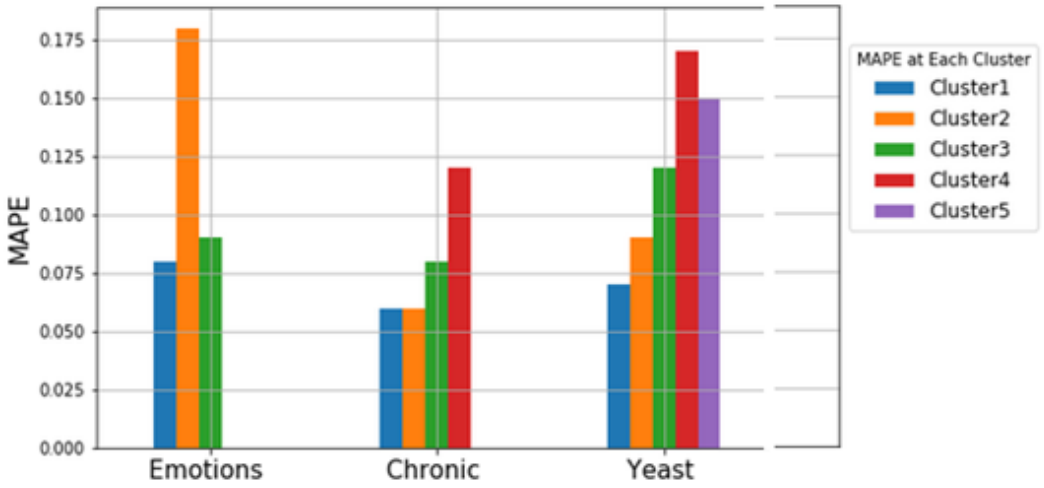


Figure 6.4: MAPE of the clustering algorithm on each cluster in each dataset

than 10% indicates highly accurate forecast,11 to 20% is a good forecast, 21 to 50% is a reasonable forecast, and 51% or more is an inaccurate forecast. Accordingly, a highly accurate forecast is found in the chronic disease dataset. The results on emotion and yeast datasets show a good prediction.

## 6.2 Identifying Subgroups of Elderly Patients using Clustering Analysis

### 6.2.1 Introduction

In order to address the specific healthcare needs of a population, it is vital to develop more effective healthcare models that can be applied for the decision-making process [344, 345]. Specifically, elderly-driven care approaches continue to become a central issue for healthcare systems and to face the different challenges associated with the frail people. To really center on the elderly, his or her specific care needs and other characteristics must be addressed. While it is practically expensive to develop healthcare models and intervention programs for every individual, programs can be proposed for subgroups of patients with similar characteristics. In this regard, clustering presents as an appropriate method for dividing an elderly population into distinct groups, each with specific needs, characteristics, or behaviors to allow care delivery and policies to be tailored for these groups [35].

In this section, clustering methods have been used for identifying subgroups of patients among the elderly with homogeneous characteristics by exploring the underlying structure of the data. Currently, the wide adoption of electronic health records in healthcare systems makes administrative data more accessible and provides opportunities to utilize real patient records for different data analytics and prediction tasks [344]. Data from administrative systems can be used to allocate patients into clusters based on, for example, medical health conditions, and analyze healthcare costs and adverse outcomes per clusters. With the growing awareness of administrative data as an asset, clustering approaches are adopted in order to identify and address the distinct healthcare profiles and priorities of different groups comprising it [346]. Data-driven clustering approaches empower healthcare systems to have a deeper understanding of elderly health needs, facilitate effective healthcare resource planning, and improving many aspects of present and future healthcare delivery [347].

In this section, the study aims to identify groups of elderly people using a cluster analysis approach based on administrative data associated with frailty syndrome. The discovered clusters can serve as a robust basis for interpreting potential correlations among patient characteristics within the context of chronic conditions, comorbidities, and other factors.

### 6.2.2 Materials and Methods

#### Data Source

Approximately 1.1 million medical records of the elderly patients aged 65 and above have been used for the clustering problem. As described in chapter 4, the data corresponds to medical claims from the regional patients' information (including hospital discharges, mortality registers, and drug prescriptions) in Piedmont, Italy. The data provides a rich source of information on the demographic, medical, and functional characteristics of the elderly population. Each health record in the dataset has several attributes with 58 input variables and six output variables. The input variables include different combinations of demographic variables (including gender, age, marital status, household composition, etc.), socioeconomic variables (such as income, education, employment, housing, etc.), and comorbidity measures, which include Charlson comorbidity index and other chronic conditions. Chronic medical conditions include diabetes, respiratory disease,

kidney disease, stroke, dementia, cancer, congestive heart failure, depression, and many others. All categorical variables with chronic conditions were represented by Boolean values, while each non-categorical variable was specified using the dummy variables. In particular, the variable 'age' was grouped into three categories using nine-year age ranges, with 65-74 used as the first category. The full list of all the input variables that have been used for the formation of clusters is shown in Appendix A.

For each subject in the data, there are one or more outputs: mortality, access to the emergency department (ED) with red code, urgent hospitalization, disability, fracture, and preventable hospitalization. These variables are described as outcomes or measurable changes in the health status of patients. The output variables were not used for the creation of clusters but for interpreting or characterizing the clusters formed through the input variables. A cluster analysis was performed in order to identify the patterns of input cases and to allow the assignment of variables into groups or clusters so that medical conditions in the same cluster are more similar to one another than to conditions from different clusters.

### Dimensionality Reduction Methods

Dimensionality reduction method was an important aspect of our clustering problem. Three dimensionality reduction methods were explored to visualize and reduce the dimensionality of the data: Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and multiple correspondence analysis (MCA).

PCA [348] is one of the most commonly used dimensionality reduction methods which aims to find the linear combinations of the multivariate data that captures a maximum amount of variance. PCA performance depends on the presence of discrete data. Binary and ordinal discrete data can affect how PCA operates and the interpretation of the results. The t-SNE is a non-linear projection method for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets [337]. This method works by taking the high dimensional space of the data and computing a probability distribution based on points that are next to each other in the n-dimensional space. Unlike PCA and t-SNE, which are best suited for continuous data, we also used MCA, which is designed for categorical data. MCA is an extension of correspondence analysis [349] for multivariate datasets. It projects a given dataset in a lower-dimensional

subspace producing two major effects: It reduces the dimensionality of the dataset, and it projects the observations on continuous space. In particular, k-means requires continuous features on comparable scales so as not to be biassed towards features with large value ranges. The orthogonal (uncorrelated) features created by MCA ensures that highly correlated variables do not dominate cluster assignments.

The examination of different methods helped us to identify the right dimensionality technique. In each dimensionality technique, an optimal number of dimensions extracted and percentages of inertia were determined by the scree plot. However, not all of them were designed to work well with this dataset. When using MCA, we observed the meaningful distribution of input and output cases that makes each cluster to vary by unique characteristics or prevalence. Whereas in using PCA or t-NSE, we noted a higher overlap of clusters' characteristics in which we cannot identify clusters with unique behaviors. This can be because PCA and t-SNE don't work well with categorical variables in contrast to MCA, which works well and designed for categorical variables. To get consistent and clinically interpretable patterns of association in medical conditions, we considered only the results of clusters obtained on the MCA subspace. In this case, all the clustering results described in this section were based on the MCA subspace only unless otherwise specified.

### Clustering Methods

For the cluster analysis, k-means (KM) clustering, model-based clustering via the Gaussian mixture model (GMM), and hierarchical density-based spatial clustering of applications with noise (HDBSCAN) were assessed and explored. KM is one of the most widely used clustering techniques. The "K" in K-means refers to the number of clusters that should be assigned by users [341, 345]. The KM clustering uses a simple iterative technique to group points in a dataset into clusters that contain similar characteristics. The algorithm is made up of two different phases. First, the centroids are selected randomly, with a fixed value of K, and the second phase is to assign each data point to the closest center [350]. Euclidean distance is mostly used to measure the distance between cluster centers and each data point.

The GMM is based on the assumption that the data are generated by a collection of models, with each cluster corresponding to a different model [351]. Each resulting cluster is represented by a parametric distribution and can be either spherical or ellipsoidal of varying sizes and variance. The

GMM attempts to find a mixture of multidimensional Gaussian probability distributions that best model any input dataset [352]. It can be used for finding clusters in the same manner as k-means; however, GMM still assumes the data is coming from a mixture of probability distributions, each representing a different cluster.

A density-based method is used to discover clusters of non-spherical shapes. In order to find clusters of arbitrary shapes, clusters are modeled as a dense region in the data space, separated by sparse regions [342]. The HDBSCAN algorithm extends DBSCAN by converting it into a hierarchical clustering algorithm and then using a technique to extract a flat clustering based on the stability of clusters [353, 125]. It means that it can cluster data points that are close together. Then, it will mark outliers that are in the lower-density areas. HDBSCAN is different from the above two clustering algorithms because it doesn't require input to determine the number of clusters.

All three clustering techniques (KM, GMM, and HDBSCAN) have been applied to the frailty data used in this thesis. GMM works in some ways, similar to KM, specifically for our problem. However, KM solutions were slightly more meaningful and interpretable for further analysis and evaluation than GMM. As a result, we only focused on the clustering solutions of K-Means and HDBSCAN, as discussed in the results section.

### Evaluation of Cluster Qualities

One of the most important and difficult activities in clustering analysis is measuring cluster qualities, which represents the objective and quantitative assessment of clustering results [323]. A common approach to evaluate the quality of clustering results involves the use of internal validity criteria [120]. Many of such measures are designed to show the compactness and separation of data. The compactness means that the data within the same cluster should be close to each other, and the separation means that the data in different clusters should be widely spaced.

In this study, we use internal validity [320] to measure the qualities of clustering results. We also used a new method for evaluating clustering results. It is based on predicting the probability of occurrence of the six outcomes in new patients of a specific cluster, based on the characteristics of outcomes/labels in the data. A more detailed procedure and measurement techniques for the proposed cluster evaluation method are provided in section 6.1.

### 6.2.3 Experimental Analysis

Analyses were stratified by gender and age groups (65-74, 75-84, and >85 years). There were no missing values, as gender, age group, chronic diseases, and others were recorded for all the samples. Descriptive statistics were employed to summarize the overall data, and all the variables were expressed as frequencies and percentages.

Clustering has been applied to the electronic health records of elderly data to group similar patients based on various input characteristics, including demographics and socioeconomic characteristics and chronic conditions. We examined the different clustering techniques as well as feature extraction methods for identifying similar groups of elderly people using data stratified by age and gender groups.

Generally, we used a two-step process: Firstly, we took our dataset and used a data dimensionality reduction technique to convert it into a lower-dimensional dataset. Then, we took that reduced dataset and applied various clustering techniques. Alternatively, we first applied our clustering method to the data and then used our dimensionality reduction technique to visualize it. However, clustering data with all of its dimensions is possible with lower-dimensional and continuous data. Since our data contains binary and multi-valued categorical variables, it is not appropriate computing distances between observations. As a result, we emphasized reducing the dimension first and then applied the clustering algorithm on the reduced data.

On the numerical components created in the MCA step, elderly patients were classified in clusters according to proximity criteria using the K-Means algorithm with random initial centroids. An optimal number of clusters was assessed according to Calinski Harabaz criteria and the elbow method [355, 356]. Both methods showed that there were three potential clusters of elderly patients that best separated the dataset.

To describe and characterize clusters, we used the input and output variables. The input variables are cluster generating features, which can be described as comorbidities, chronic disease, demographic and socioeconomic variables. We also used the pre-defined output variables to evaluate and interpret (but not to generate) clusters: mortality, access to ED with red code, urgent hospitalization, and disability, fracture, and preventable hospitalization. Once the clusters were formed using input variables, the

prevalence of the outcome variables in each cluster were assessed and analyzed. In order to facilitate interpretation, the frequencies and percentage of each input and output features in each cluster were calculated. Cluster characterization and prevalence of variables in clusters are illustrated using bar graphs.

To assess the internal cluster quality, we applied silhouette analysis on training and test data. Using the silhouette coefficient, the degree of separation between clusters, as well as the consistency of silhouette scores between the training and testing, were determined. The dataset was randomly split into a training set (70%) and a test set (30%). First, the clustering algorithm has been applied to the training data to create potential clusters, and the test data has been used as a sample of new patients to be assigned to the existing clusters. In this case, the quality of the model to assign new patients into their nearest cluster as well as to predict the probability of occurrence of the six adverse outcomes in clusters were measured.

The majority of clustering analyses were carried out using different libraries in Python, and we implemented dimensionality reduction techniques and clustering methods as well as plotting and interpreting the results. RStudio was also used for some specific tasks, such as analyzing the dimensions of the data.

### 6.2.4 Results and discussions

The data was composed of 1,095,613 elderly patients aged 65 and above years. Both male and female patients aged 65 and older years were included in the cluster analysis, each composed of three age groups. Women outnumber men among the elderly people aged between 65 to 74 years (47% men versus 53% women), 75-84 years (44% men versus 56% women) and a predominance of women among the most elderly ($\geq$ 85 years) (33% men versus 67% women).

Both HDBSCAN and K-Means algorithms identified three clusters across all age and gender groups. An example of a two-dimensional representation of clustering results for men across all age groups is shown in Figure 6.5 for K-Means and Figure 6.6 for HDBSCAN. The shapes of the colors indicate the size of the three clusters and their cluster memberships.

In clustering results via HDBSCAN, we observed clusters that are well-separated and have stronger stability over the different portions of the dataset as opposed to overlapping clusters that are created through k-means

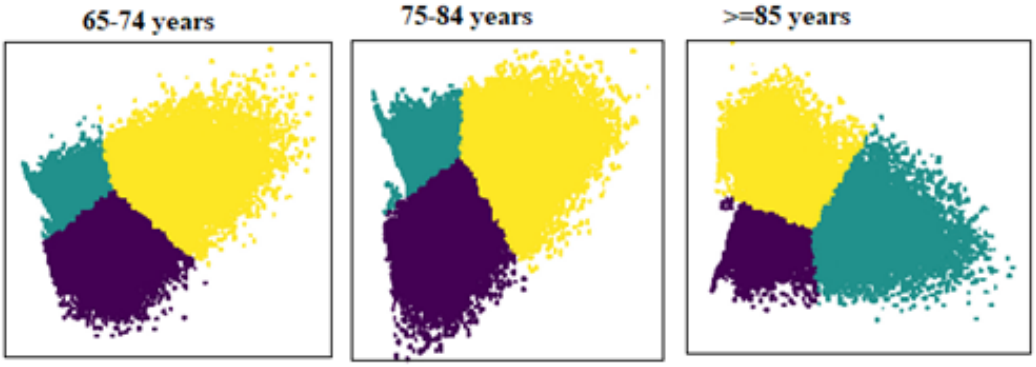


Figure 6.5: Visualizations of clustering experiments in 2D scatter plots using the K-Means algorithm for men with all age groups.

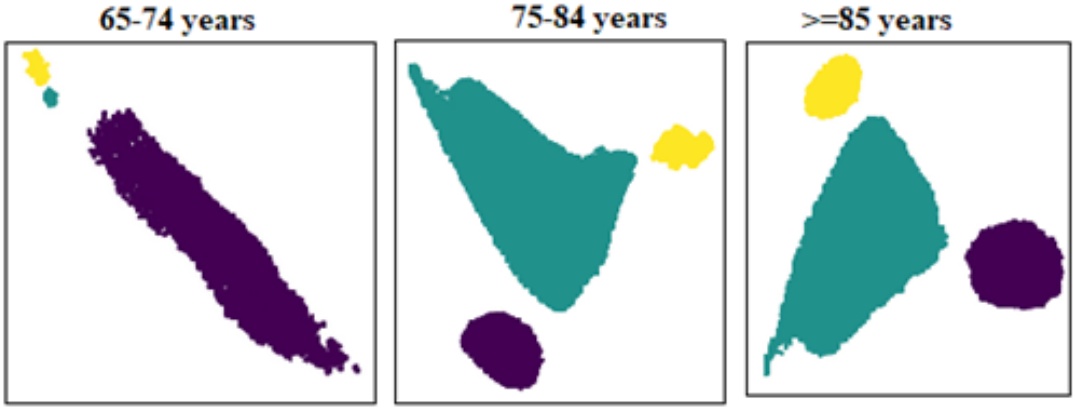


Figure 6.6: Visualizations of clustering experiments in 2D scatter plots using HDBSCAN for men with all age groups.

clustering. However, in the context of our frailty problem, HDBSCAN classifies and removes most of the frail subjects as noise, which may lose efficiency in grouping elderly patients into plausible clusters. It has also shown a high probability of getting one larger cluster with the creation of other very small outlying clusters. HDBSCAN also scored the lowest average silhouette coefficient. However, the clusters obtained using the k-means algorithm had a higher average silhouette coefficient, as shown in Table 6.2. In this experiment, the silhouette measure obtained for HDBSCAN clusters is between 0.47 and 0.71. In KM, the silhouette measure obtained for clusters is found between is 0.76 and 1.0 for all age and gender groups, demonstrating

a strong clustering structure was found. Furthermore, almost similar silhouette scores were obtained in the training and testing datasets, which shows the concordant or consistency of clusters between the training and testing datasets with respect to the input characteristics and other covariates.

Table 6.2: Average silhouette score of clusters obtained using the k-means

| Gender | Age Group | Silhouette Score | |
|---|---|---|---|
| | | Training data | Testing data |
| Female | 65-74 | 0.768 | 0.767 |
| | 75-84 | **0.802** | 0.802 |
| | 85+ | 0.773 | 0.772 |
| Male | 65-74 | 0.791 | 0.79 |
| | 75-84 | 0.802 | 0.801 |
| | 85+ | **0.762** | 0.76 |

The descriptive statistics of clusters identified through K-Means is shown in Table 6.3 for women and men in all age groups. For simple comparison and interpretation, the clusters were sorted from smallest to largest based on the number of samples. The first cluster had the smallest percentage of the sample, both women and men: 8% and 10 % of those aged 65–74 years, 11% and 15% of those aged 75-84, 15% and 18% of those aged 85 and above respectively. On the other hand, the largest percentage of patients was observed in the third cluster in all age groups of men and women: For age groups 65-74, 75-84 and 85+, there was 50%, 47%, and 48% of women patients in the third cluster, respectively.

Table 6.3: Descriptive statistics of clusters, stratified by gender and age group

| **Clusters** | **65-74 years** | | **75-84 years** | | **85 years** | |
|---|---|---|---|---|---|---|
| | Female (n=265251) | Male (n=239386) | Female (n=230872) | Male (n=178104) | Female (n=122385) | Male (n=59615) |
| **0** | 20311<br>8% | 24721<br>10% | 26013<br>11% | 26272<br>15% | 17882<br>15% | 10616<br>18% |
| **1** | 111387<br>42% | 93450<br>39% | 95374<br>41% | 67190<br>38% | 45725<br>37% | 19679<br>33% |
| **2** | 133553<br>50% | 121215<br>51% | 109485<br>47% | 84642<br>48% | 58778<br>48% | 29320<br>49% |

Differences in characteristics between clusters were compared according to the input and output variables, using Pearson Chi-square tests. The significance level was set at $\alpha$ =0.05, and all tests were two-tailed. The input variables are cluster membership or cluster generating features which are involved in the process of clusters formation, while the output variables (mortality, disability, fracture, urgent hospitalization, preventable hospitalization, and access to emergency room visits) are not used in cluster generation, but they are used for assessing, evaluating and interpreting clusters. The clusters' difference in each age groups of both genders were reviewed and assessed to identify whether they were statistically different with respect to the output variables. All input variables differed statistically significant between clusters (all P-values $< 0.00$ ) in all age groups. The output variables were also found to differ significantly across clusters (P-values $<$ 0.00). Table 6.4 presents an example of the statistical test results among the three clusters with respect to the six output variables for women aged 65-74.

Table 6.4: Comparison of output variables across the three clusters for women aged 65-74.

| Output Variable | Code (0=no, 1=yes) | Cluster 0 (N=17641) | | Cluster 1 (N=115345) | | Cluster 2 (N=85725) | | P-Value* |
|---|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % | |
| Mortality | 0 | 19364 | 95.34 | 110524 | 99.2 | 133099 | 99.7 | <0.0 |
| | 1 | 947 | 4.66 | 863 | 0.77 | 454 | 0.34 | |
| Access to ED with red code | 0 | 20131 | 99.11 | 111048 | 99.7 | 133350 | 99.9 | <0.0 |
| | 1 | 180 | 0.89 | 339 | 0.3 | 203 | 0.15 | |
| Urgent Hospitalization | 0 | 19228 | 94.67 | 109149 | 98 | 132266 | 99 | <0.0 |
| | 1 | 1083 | 5.33 | 2238 | 2.01 | 1287 | 0.96 | |
| Disability | 0 | 19259 | 94.82 | 110017 | 98.8 | 133112 | 99.7 | <0.0 |
| | 1 | 1052 | 5.18 | 1370 | 1.23 | 441 | 0.33 | |
| Fracture | 0 | 20169 | 99.3 | 111127 | 99.8 | 133382 | 99.9 | <0.0 |
| | 1 | 142 | 0.7 | 260 | 0.23 | 171 | 0.13 | |
| Preventable Hospitalization | 0 | 19780 | 97.39 | 110523 | 99.2 | 133253 | 99.8 | <0.0 |
| | 1 | 531 | 2.61 | 864 | 0.78 | 300 | 0.22 | |

### Prevalence of Chronic Disease and Hospitalization

Figure 6.7 presents the prevalence of the most common chronic disease and

hospitalization conditions of men and women clusters for the age group 75-84. Almost, in all clusters, we observed the heterogeneity in the occurrence of chronic disease and hospitalization conditions among the three clusters through all age and gender groups. Among the three clusters formed in all age groups, the third cluster is the one that contains the largest cluster size with the lowest prevalence of hospitalization and chronic conditions in patients as opposed to the first cluster, which had the largest prevalence. In all age groups, the first cluster (cluster 0) has been characterized by the largest occurrence of hospitalization variables (number of urgent hospitalization, non-traumatic hospitalization, emergency admissions with red code and total hospitalization) with the urgent hospitalization being the highest in clusters of older age groups (85+ years). The overall prevalence of total hospitalization was significantly higher in men than in women in all clusters of all age groups, whereas some diseases such as mental disease, depression, and femur fracture are significantly more prevalent in women than men across all clusters of age groups. Comparing the first cluster across all age groups, we noticed substantial heterogeneity in the composition of urgent hospitalization with a proportion of 41%, 57%,75% for women and 46%, 57%, 72% for men with age groups 65-74, 75-84 and >85, respectively. This shows intuitively that the oldest age group (85+) had the highest probability of being hospitalized than the lower age groups in both gender groups.

The second cluster (cluster 1) can be characterized by the predominant occurrence of hypertension and cardiovascular disease both in men and women. This cluster had the smallest occurrence of hospitalizations and emergency admissions. This cluster also had a significant prevalence of diabetes, respiratory disease, anemia, depression, blood disease, nerve disease, and so on. In both men and women, there was a nearly similar prevalence of clinical conditions in the second cluster. In the remaining cluster, there is always the least number of chronic diseases and other input cases.

The third cluster contained none of the emergency department visits with red code and urgent hospitalization in all age groups. The prevalence of input cases such as the number of urgent hospitalization, number of emergency admissions with red code, femur fracture, circulatory system disease, increases significantly with aging for men and women, whereas some other diseases such as cancer, hypertension, and diabetes are nearly similar across all age groups.

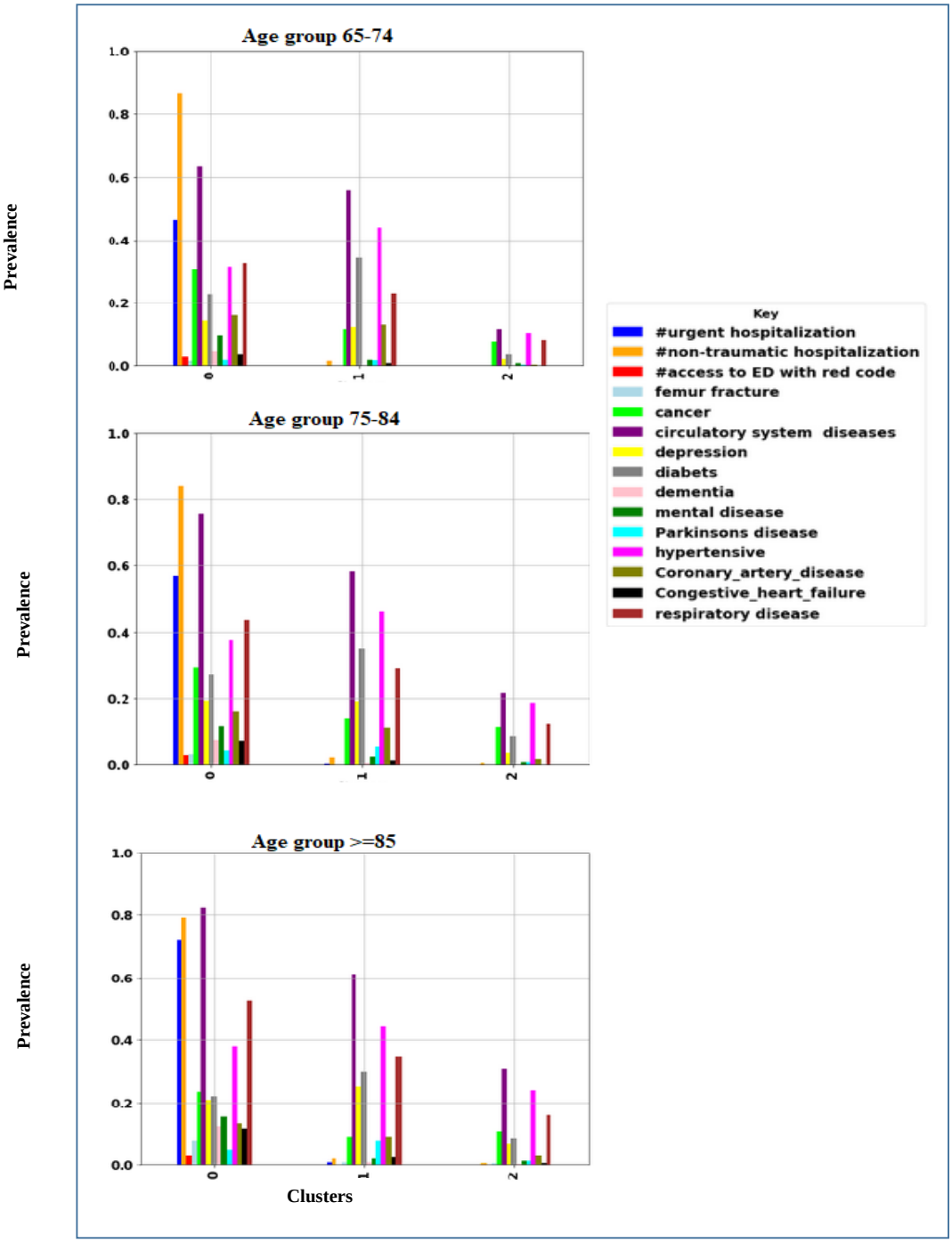


Figure 6.7: The prevalence of input variables in the three clusters of different age groups

### Probability of Outcomes in Clusters

The probability of occurrence of mortality, emergency admissions with red code, urgent hospitalization, and disability, preventable hospitalization, and fracture outcomes for all clusters across all age and gender groups are shown in Figure 6.8. Once the clusters are formed, the probability that a problem will occur, given that there is a cluster with negative and positive cases, mathematically written as *'P(Problem | Cluster)'* can be estimated. The first cluster had the highest probability of occurrence of the six outcomes, followed by the second and third clusters in all age groups. The probability of occurrence of the mortality and urgent hospitalization is a little higher in men than women, while disability and fracture are more prevalent in women than men. As shown in the plots (Figure 6.8), it is intuitive to mention that the probability of occurrence of all outcomes (mortality, emergency admissions with red code, urgent hospitalization, and disability, preventable hospitalization and fracture) increases from the younger age groups (65-74) to the oldest age groups (85+) in each of the clusters. For example, the probability of a woman having an urgent hospitalization is 5% for age groups 65-74 years, 8% for 75-84 years, and 9% for 85+ years.

For age groups 65-74, both genders, the most prevalent outcome is urgent hospitalization followed by mortality. For age groups 75-84, the most prevalent outcome is a disability for women and mortality for men. In the oldest age groups (85+years), mortality is the predominant,accounting an average overall rate of 16% for men and 13% for women.

Generally, the number of patients who were dead, hospitalized, disabled, admitted with red code, and /or fractured are the most dominant in the first cluster. The subgroups in both genders can be roughly labeled according to the dominant input and output variables assigned to patients as follows: Cluster 0 – Frail , Cluster 1 – Pre- frail, Cluster 2 – non-frail. In this regard, early intervention-oriented systems should gain priority attention for elderly patients in the first cluster (cluster 0). This cluster requires urgent interventions addressing some of the most common problems encountered in older patients.

Once the probability of outcomes in training data is determined, the probability of urgent hospitalization, disability, fracture, emergency admission, or death for next year in adults aged over 65 in the testing data can be estimated. In each cluster for all age groups, the probability of occurrence of outcomes in the test data was calculated and predicted.

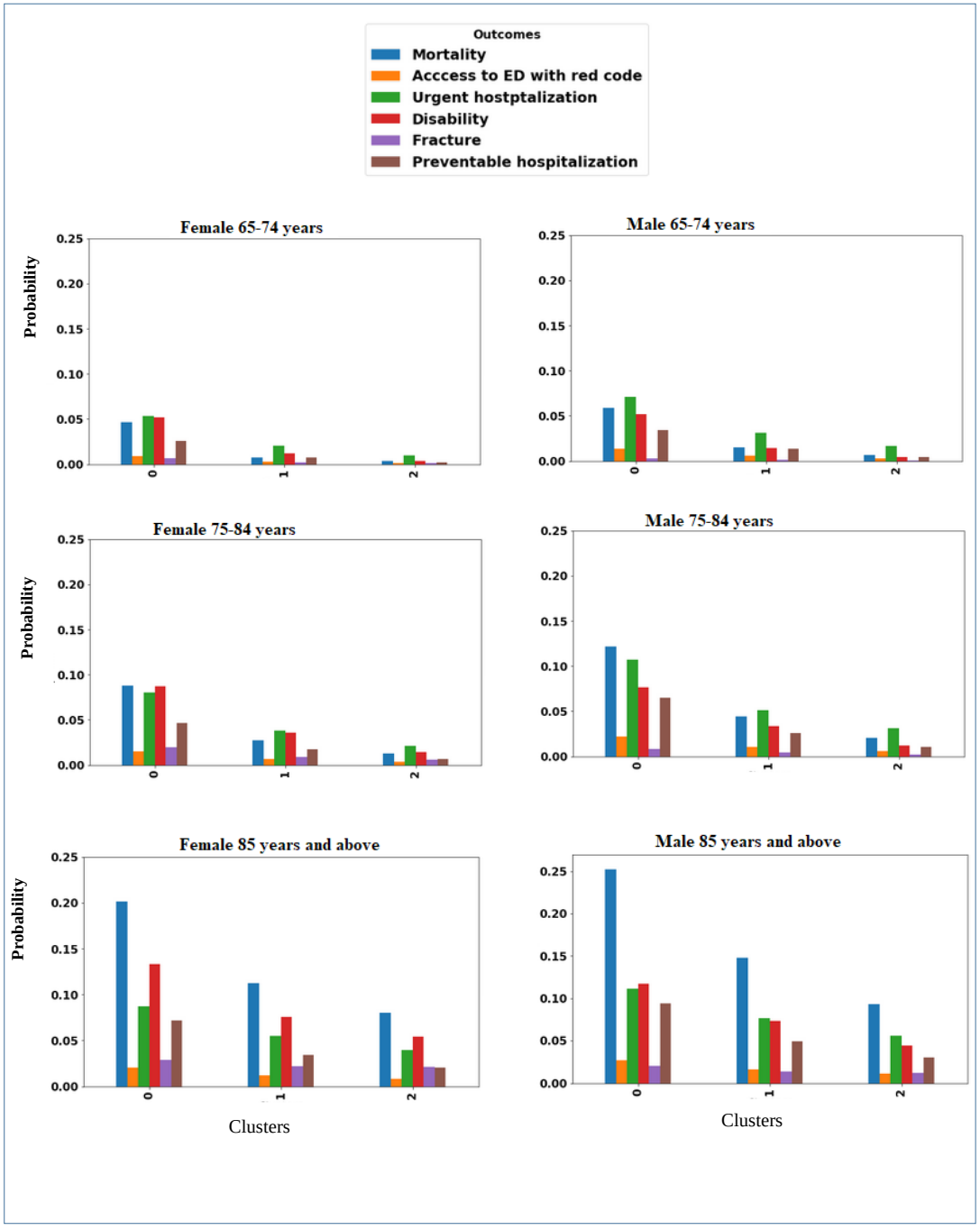


Figure 6.8: Probability of occurrence of the six outcomes in clusters, stratified by age and gender groups

The prediction capabilities were evaluated by using cross-validation procedure, as proposed in section 6.1. Tabel 6.5 presents the quality of each cluster measured interms of MAPE and RMSE.

Table 6.5: Cluster quality measures in terms of MAPE and RMSE

| **Gender** | **Age Group** | **MAPE in %** | | | **RMSE** | | |
|---|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 1 | Cluster 2 | Cluster 3 |
| | 65-74 | 2.73 | 2.96 | 1.99 | 0.00177 | 0.00071 | 0.00489 |
| Female | 75-84 | 5.45 | 6.75 | 1.4 | 0.00897 | 0.0095 | 0.00305 |
| | >=85 | 1.91 | 0.64 | 1.54 | 0.00643 | 0.00495 | 0.00659 |
| | 65-74 | 1.12 | 1.75 | 1 | 0.00092 | 0.00146 | 0.00403 |
| Male | 75-84 | 2.88 | 2.44 | 0.86 | 0.00396 | 0.0044 | 0.00389 |
| | >=85 | 3.29 | 2 | 2.75 | 0.0083 | 0.01006 | 0.01404 |

Clusters were also evaluated on the basis of results obtained from the training dataset. In order to test the reproducibility of the resulting clusters on the test dataset, we trained one of the supervised learning classifier (e.g., random forest classifier) [354], using the labels acquired from the clustering process, and validate the model on the remaining subset. To evaluate the robustness of the clustering results, we repeated the experiment (MCA/k-means clustering) to the test set and compared the concordance of the acquired clusters with the cluster labels predicted by the random forest classifier (RF).

The Random Forest (RF) was trained on 70% of the clustering training data with the K-Means (KM) cluster labels as the predicted classes using the 56 features. For example, for women 65-74 years old, the accuracy of RF on the test part of the data was able to reach 96% accuracy in classifying unseen patients (test data) to their respective cluster subgroups. The characteristics of the three clusters derived from the test dataset were comparable to those of the training dataset with regards to variables included in the analysis, indicating good reproducibility of the clusters in test data. The same number of 56 variables were also used as the input to the MCA/K-Means to predict the cluster membership for test data. The similarity of cluster assignments (prediction of cluster membership) by the RF and by the MCA/K-Means was calculated using the Jaccard index (percentage of patients overlapping in the same cluster between the two solutions).We found very good agreement between the two approaches with 96% Jaccard index,

indicating that most patients are allocated to the same clusters in both the RF and MCA/K-Means approaches. An example graphical representation of the RF and MCA/K-Means outputs is shown in Figure 6.9 for women aged 65-74.

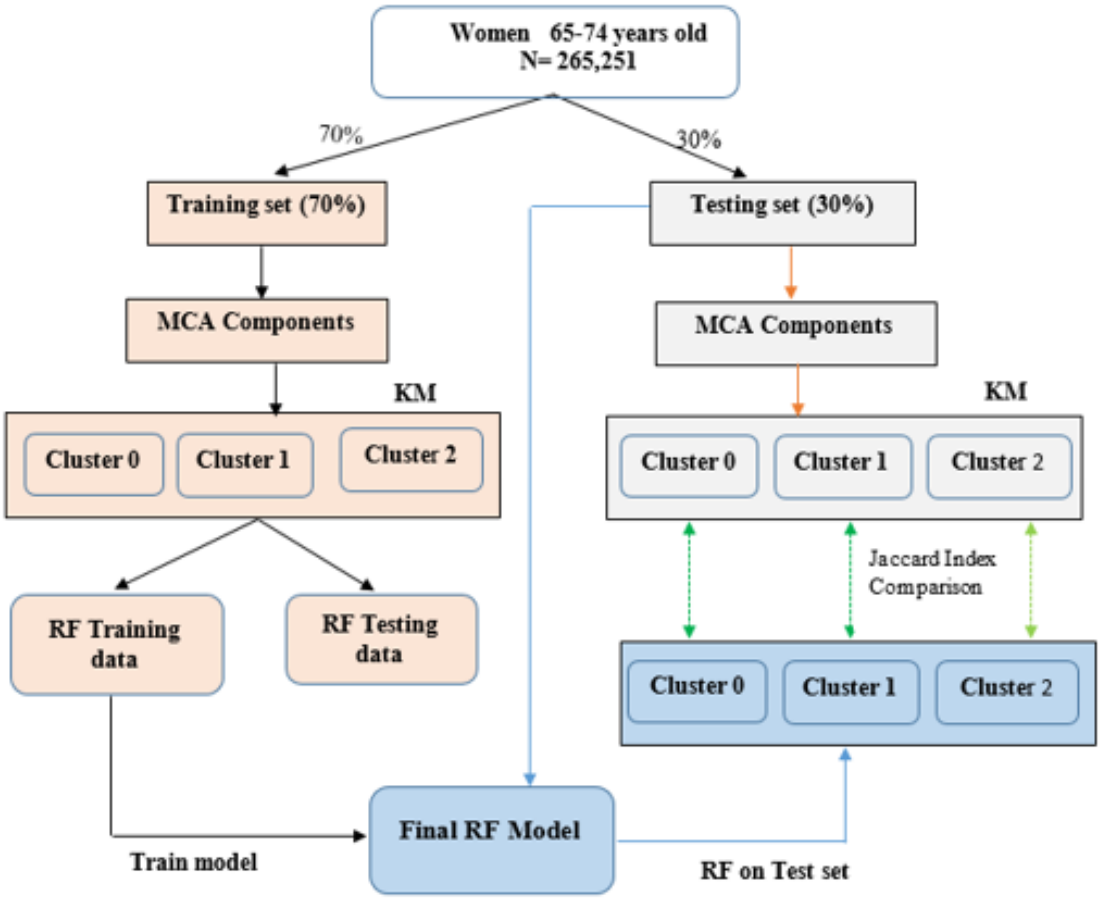


Figure 6.9: Experimental evaluation of clusters on predicting cluster membership for new patient cases using a combined approach of random forest classifier and K-Means clustering.

## 6.3 Conclusions

The first part of this chapter focuses on evaluating the quality of clustering algorithms, which is an important and challenging part of the clustering task. In this study, the k-fold cross-validation procedure was adapted to the task of evaluating the quality of the clustering algorithms, that is, measuring the ability of these algorithms to predict cluster membership for new

data. A new clustering validity index was proposed to measure the effectiveness of the clustering algorithm through the use of root mean squared error (RMSE) and mean absolute percentage error (MAPE) values. The index was developed using the probability information obtained from several labels of multi-label data. This measure is useful for evaluating clusterings, which can be used for estimating the probability of the occurrence of the labels. For example, patients can be grouped into several clusters, and the occurrence of diseases can be studied separately in each group. The results presented in this chapter show that the proposed method works well for evaluating the quality of clusters obtained using the k-means algorithm. Combining the proposed method with other, for example, density-based, clustering algorithms require solving additional problems such as finding an effective way of assigning new data points to previously discovered clusters. Therefore, combining the proposed method with such clustering algorithms was left as further work.

In the second part of this chapter, we applied a clustering analysis to the administrative data of elderly people routinely collected primarily from hospital discharges,drug prescriptions, etc. This data-driven analysis has identified three distinct, plausible clusters of elderly patients and the relationship between variables within each cluster. The first cluster is composed of the highest-burden of diseases and outcomes, which requires urgent action as a key priority to maintain health and avoid future costs in this cluster. In general, the findings provide the basis for further research into the consequences of frailty clustering in terms of outcomes and have particular significance for frailty prevention plans, clinical practice, and the planning of follow-up services for older people.

# Chapter 7

# Conclusions and Future Work

This chapter provides a summary and recommendations for future research work

## 7.1 Summary of the Thesis

Currently, several countries around the world face problems associated with an increasingly aging population, which drives significant challenges for their health services and societies. One of the common challenges is the growth of older adults who have a frail health condition or are at a greater risk of progressing frailty conditions. Frailty is a complex phenomenon as the target population of frail elderly adults may have diverse and complex needs. It has been a major threat to older peoples' wellbeing along with high service expenditures. Therefore, it should be detected at an early stage to give options for proactive care, in order to help reduce or delay functional decline, prevent hospitalization, and maintain the wellbeing and independent life of the elderly. For the detection of frailty, various frailty scores based on different concepts of frailty have been developed. However, so far, none of them is recognized as the "gold standard", and the debate is centered on whether frailty should be defined solely in terms of biomedical factors or whether other factors should also be considered.

In different clinical and social settings where the care of the elderly is a priority, early detection initiatives should consider the current views on frailty that integrates various factors including, clinical, cognitive, psychological and socioeconomic factors, in order to reflect the multidimensional impairments and consequences that are intrinsic to the frailty syndrome

[9]. In current clinical settings, health administrative databases are important source of information to assess frailty from various dimensions. Such databases are also available for health services planning and delivery, surveillance of health status population, or various researches [24]. An administrative health database with large and heterogeneous dimensions requires a better analytical method that could capture the interaction of various components. Machine learning represents a powerful approach to process the complex nonlinear relationships between various factors and yield more stable predictions.

This thesis is focused on the analysis of the administrative health database of the elderly population using machine learning approaches, both the supervised and unsupervised learning paradigms. In particular, the presented work has explored the application of a classification task for detecting and predicting frailty, and clustering analysis for the management of frail older adults. In chapter 4, we have investigated the use of various standard (i.e., single-label) classification models to identify various adverse health outcomes associated with frailty (fracture, disability, medical emergency admission at the emergency department, urgent hospitalization, preventable hospitalization, and premature mortality) from a routinely collected medical administrative dataset, including 1,095,613 older adults aged 65 years and above from Piedmont, a region in the North-West of Italy. Important preprocessing procedures, such as feature selection and reducing the imbalanced nature of the data, were performed before building machine learning models. As we confirmed from the results of the experiments, the prediction performance of machine learning models significantly varies from one problem (i.e., from one outcome associated with frailty) to another in terms of different evaluation measures and the dataset used in the study is better at predicting mortality than predicting the other outcomes associated with frailty.

Although the explored models have shown a strong predictive ability to estimate the risk of a single outcome associated with frailty syndrome, they are not well aligned to handle multiple labels simultaneously, as the data originally contains multiple negative health outcomes. Moreover, current studies on single-label classification for complex multi-label datasets fail to handle new approaches of improving performance through exploiting label correlations. Thus, chapter 5 is focused on detecting more than one

adverse outcomes concurrently using the multi-label learning method. Various multi-label learning algorithms have been employed and compared for simultaneously predicting the six adverse outcomes associated with frailty. One of the main challenges in multi-label learning is the joint occurrence of highly imbalanced labels in the same data patterns. A hybrid resampling approach, named ML-TLSMOTE, was proposed to reduce the imbalanced label distribution and improve the performance of MLC algorithms.

In chapter 6, the problem of identifying the profiles of frail elderly people was tackled through cluster analysis. The aim was to divide the elderly patients into distinct groups, each with specific needs, characteristics, or behaviors to allow healthcare delivery and policies to be tailored for these groups. This thesis has identified three distinct, plausible clusters of elderly patients across all age and gender groups, and that it provides a unique opportunity for investigating the co-occurrence of multiple conditions, the rate of comorbidities and the prevalence of adverse outcomes in each cluster of elderly patients. The results can help us to know those clusters composed of patients with frailty, and therefore, require urgent action and priority to maintain health and avoid future costs. One of the main issues related to the clustering task is the proper validation of the clustering results, which is usually reported as one of the most difficult and frustrating steps of cluster analysis [117]. In this thesis, we propose a new cluster validity index that uses the information from multiple labels to evaluate the quality of clustering algorithms. The study validates the proposal through the cross-validation analysis of some challenging multi-label datasets. This approach was motivated by medical applications in which we would like to assess the probability of various health problems in different patient groups. For example, the labels for the frailty dataset are mortality, hospitalization, disability, and fracture, as indicated in section 4. Once the clusters are formed, the probabilities of the occurrence of these labels, i.e., mortality, hospitalization, disability, and fracture are estimated in each cluster and compared between the training set and the test set. The aim is to measure how well we can predict the probabilities of these three outcomes in new patients (i.e., in the test data) based on their membership in the training clusters.

Finally, it is important to highlight that the various approaches proposed in this thesis were demonstrated and validated using an administrative health

database of older adults, and the results clearly showed that machine learning models are particularly relevant for early intervention and prevention of frailty syndrome in the elderly population. Currently, this is the first to attempt to develop frailty predictive models using a large population of routinely collected socio-clinical administrative data with multiple simultaneous outcomes, and the first to attempt to do so using machine learning, in particular using multi-label learning methods. Although traditional statistical modeling is frequently used to identify frailty and its risk factors, machine learning has become more indispensable for solving more complex problems and provides highly useful information for frailty risk prediction and for informing courses of treatment. Such sophisticated predictive knowledge is often of great importance to physicians, health specialists, policymakers, or other professionals, who may then advice to screen high-risk individuals or early interventions to prevent adverse health outcomes from happening in patients. Furthermore, modern machine learning techniques, unlike the classical methods, have greater flexibility to capture non-linear complex interactions among a large number of variables in massive quantities of data, making them more suitable for big steps in diagnostics and prediction. Therefore, an increased integration of machine learning into everyday medical applications could improve the efficiency of treatments and lower costs in various ways. Despite such in-depth applications of machine learning in healthcare and other domains, some limitations still exist, including the data acquisition and representation issues [357], the degree of interpretability for predictive power [358], and the deployment issues [359].

## 7.2 Future Research Directions

The final consideration of this thesis is to show some future directions of research related to the application of machine learning techniques in a frail older population. While the investigation of this thesis have shown promising results with respect to the detection and management of frailty as well as some approaches for enhancing the quality of predictive models, there is still much work that requires further investigation and research.

The thesis used an administrative health database of the elderly for analyzing and predicting the risk of frailty syndrome (fracture, disability, urgent hospitalization, and mortality). With appropriate modifications through an

extended tuning process and by investigating various socioclinical and administrative databases at the national level, the predictive models might be applicable in an international setting. Furthermore, the frailty phenotypic information such as grip strength and gait speed can be integrated with administrative databases in order to develop a comprehensive model with improved predictive ability.

One of the issues that pose a challenge in building a predictive model was that the data suffer from the imbalanced class distribution. To address the imbalanced problem in this administrative data, we applied commonly used resampling methods. These approaches are model-independent preprocessing techniques that can be applied to any machine learning algorithm. However, they may lead to loss of important information or the introduction of meaningless new samples. Therefore, alternative approaches such as cost-sensitive learning [238] can be integrated with the new emerging machine learning methods such as deep learning, resulting in robust and efficient learners without the perturbation of the original data.

Multi-label feature selection is a powerful tool for a high-dimensional problem in order to improve the performance of multi-label classifiers. In this thesis, the single-label based feature selection methods such as chi-square and information gain have been tried in the multi-label scenario to enhance the performance of classifiers. However, the existing feature selection methods are unable to take both computational complexity and label correlation into consideration. Moreover, since each label might be decided by some specific features of its own, the problems of feature selection are often addressed independently. Therefore, further study can be important to perform joint feature selection in a multi-label classification that can learn both shared features and label specific features by considering pairwise label correlations.

In developing a multi-label frailty risk prediction model, this thesis presents a new resampling algorithm for imbalanced multi-label classification problem named ML-TLSMOTE (multi-label resampling using SMOTE followed by Tomek link), which is based on the existing hybrid resampling approaches. ML-TLSMOTE is an effective technique that reduces the imbalanced label distributions and improves performance in a multi-label data when compared to the independent resampling techniques such as SMOTE or Tomek link. The ML-TLSMOTE algorithm was evaluated considering

the specific problem of the administrative database of older adults. In the future, we need to evaluate our proposed algorithm using other benchmark datasets.

The thesis was also focused on the application of the clustering problem for identifying new frailty classes using administrative health database. Two essential aspects of this clustering problem were: (1) to estimate the number of clusters using input features in the dataset; and to allocate the output features to these clusters so that we can compute the probability of occurrence of adverse health outcomes in each cluster, and (2) assess the prediction strength of the clustering model to predict cluster labels for new samples. For the second problem, a new cluster validation criteria was proposed, named CVIM, using the information from multiple labels of multi-label data. The results show that the CVIM works well for evaluating partitioning based clustering results (e.g., k-means). Combining the CVIM approach with others, for example, density-based clustering algorithms require solving additional problems such as finding an effective way of assigning new data points to previously discovered clusters.

With the growing number of an older population and its associated challenges in healthcare demand and social care, the management or reduction of healthcare costs while improving the quality of service is a priority. In order to address these challenges, remote monitoring of older person's health can be used to identify pre-frailty and frailty conditions, therefore, enabling early intervention and reduce hospital admissions. Advanced machine learning techniques can be used for the detection of elderly health outcomes with multiple data sources such as wearable devices, sensors, and video cameras. Thus, the development of a smart healthcare monitoring system, which is capable of observing elderly people remotely, can be considered important as future development.

# Appendices

# A Description of input variables

| Variable | Category | Code | Count | Percent (%) |
|---|---|---|---|---|
| Age | 65-69 | 0 | 273,389 | 24.95 |
| | 70-74 | 1 | 231,248 | 21.11 |
| | 75-79 | 2 | 233,122 | 21.28 |
| | 80-84 | 3 | 175,854 | 16.05 |
| | 85-89 | 4 | 117,674 | 10.74 |
| | >=90 | 5 | 64,326 | 5.87 |
| Citizenship | PSA* | 0 | 1,076,375 | 98.24 |
| | PFPM** | 1 | 19,238 | 1.76 |
| Number of urgent hospitalization | 0 | 0 | 1,024,739 | 93.53 |
| | 1 or 2 | 1 | 67,523 | 6.16 |
| | 3 or more | 2 | 3,351 | 0.31 |
| Number of non-traumatic hospitalizations | 0 | 0 | 987,152 | 90.10 |
| | 1 or 2 | 1 | 99,576 | 9.09 |
| | 3 or more | 2 | 8,885 | 0.81 |
| Number of total hospitalizations | 0 | 0 | 977,028 | 89.18 |
| | 1 or 2 | 1 | 107,881 | 9.85 |
| | 3 or more | 2 | 10,704 | 0.98 |
| Charlson index [370] | Value index 0 | 0 | 1,019,406 | 93.04 |
| | Index 1-2 | 1 | 61,717 | 5.63 |
| | Index 3 - 5 | 2 | 10,160 | 0.93 |
| | value index 6 or higher | 3 | 4,330 | 0.40 |

| | | | | |
|---|---|---|---|---|
| Home-based care | no | 0 | 1,084,324 | 98.97 |
| | yes | 1 | 11,289 | 1.03 |
| Income | no | 0 | 251,424 | 22.95 |
| | yes | 1 | 844,189 | 77.05 |
| Invalidity | no | 0 | 959,791 | 87.60 |
| | yes | 1 | 135,822 | 12.40 |
| Poly prescriptions (number of drugs prescribed) | from 0 to 5 drugs | 0 | 559,731 | 51.09 |
| | from 6 to 10 drugs | 1 | 366,355 | 33.44 |
| | more than 11 drugs | 2 | 169,527 | 15.47 |
| Number of different types of drugs prescribed ( First three digits of ATC code) | 0 | 0 | 82,879 | 7.56 |
| | from 1 to 5 | 1 | 566,187 | 51.68 |
| | from 6 to 14 | 2 | 439,599 | 40.12 |
| | 15 or more | 3 | 6,948 | 0.63 |
| emergency department visits with white code | no access | 0 | 1,062,489 | 96.98 |
| | at least one access | 1 | 33,124 | 3.02 |
| emergency department visits with green code | no access | 0 | 879,634 | 80.29 |
| | 1 or 2 accesses | 1 | 194,310 | 17.74 |
| | 3 or more accesses | 2 | 21,669 | 1.98 |
| emergency department visits with yellow code | no access | 0 | 1,039,685 | 94.90 |
| | 1 or 2 accesses | 1 | 54,426 | 4.97 |
| | 3 or more accesses | 2 | 1,502 | 0.14 |
| emergency department visits with red code | no access | 0 | 1,092,082 | 99.68 |
| | at least one access | 1 | 3,531 | 0.32 |
| Housing condition | privately owned | 1 | 890,222 | 81.25 |
| | renting | 2 | 157,150 | 14.34 |
| | other | 3 | 48,241 | 4.40 |

| | | | | |
|---|---|---|---|---|
| Marital status | single | 1 | 81,492 | 7.44 |
| | married | 2 | 697,589 | 63.67 |
| | widower | 3 | 251,423 | 22.95 |
| | divorced | 4 | 65,109 | 5.94 |
| Level of education | degree or superior average | 1 | 188,799 | 17.23 |
| | lower average or professional qualification | 3 | 325,242 | 29.69 |
| | elementary or without study title | 4 | 581,572 | 53.08 |
| Work status | employed | 1 | 95,188 | 8.69 |
| | housewife | 3 | 102,523 | 9.36 |
| | withdrawn from work | 5 | 851,202 | 77.69 |
| | Other (unemployed, student, etc.) | 7 | 46,700 | 4.26 |
| Home living status | noncrowded | 1 | 1,020,696 | 93.16 |
| | crowded | 2 | 74,917 | 6.84 |
| Type of family | an elderly couple (both over 65) without children | 1 | 375,189 | 34.24 |
| | only with children | 2 | 63,561 | 5.80 |
| | only without children, single or widower | 3 | 360,051 | 32.86 |
| | other cohabitations | 4 | 296,812 | 27.09 |
| Disability | no | 0 | 950,048 | 86.71 |
| | yes | 1 | 145,565 | 13.29 |
| Femur Fracture | no | 0 | 1,040,449 | 94.97 |
| | yes | 1 | 55,164 | 5.04 |
| Depression | no | 0 | 929,835 | 84.87 |
| | yes | 1 | 165,778 | 15.13 |
| Diabetes | no | 0 | 914,357 | 83.46 |
| | yes | 1 | 181,256 | 16.54 |
| Arthropathy | no | 0 | 1,046,700 | 95.54 |
| | yes | 1 | 48,913 | 4.46 |

| | | | | |
|---|---|---|---|---|
| Parkinson's disease | no | 0 | 1,074,446 | 98.07 |
| | yes | 1 | 21,167 | 1.93 |
| Epilepsy | no | 0 | 1,033,834 | 94.36 |
| | yes | 1 | 61,779 | 5.64 |
| Anaemia | no | 0 | 1,018,977 | 93.01 |
| | yes | 1 | 76,636 | 6.99 |
| Hypertensive | no | 0 | 782,954 | 71.46 |
| | yes | 1 | 312,659 | 28.54 |
| Glaucoma | no | 0 | 1,016,227 | 92.75 |
| | yes | 1 | 79,386 | 7.25 |
| Mental disease | no | 0 | 1,060,794 | 96.82 |
| | yes | 1 | 34,819 | 3.18 |
| Cancer | no | 0 | 969,426 | 88.48 |
| | yes | 1 | 126,187 | 11.52 |
| Thyroid | no | 0 | 992,007 | 90.54 |
| | yes | 1 | 103,606 | 9.46 |
| Dementia | no | 0 | 1,080,388 | 98.61 |
| | yes | 1 | 15,225 | 1.39 |
| Coronary artery disease | no | 0 | 1,048,234 | 95.68 |
| | yes | 1 | 47,379 | 4.32 |
| Congestive heart failure | no | 0 | 1,082,722 | 98.82 |
| | yes | 1 | 12,891 | 1.18 |
| Stroke | no | 0 | 1,067,443 | 97.43 |
| | yes | 1 | 28,170 | 2.57 |
| Kidney failure | no | 0 | 1,084,182 | 98.96 |
| | yes | 1 | 11,431 | 1.04 |
| Hypercholesterolemia | no | 0 | 1,067,678 | 97.45 |
| | yes | 1 | 27,935 | 2.55 |
| Atrial fibrillation | no | 0 | 1,044,590 | 95.34 |
| | yes | 1 | 51,023 | 4.66 |
| neck fracture | no | 0 | 1,084,155 | 98.95 |
| | yes | 1 | 11,458 | 1.05 |
| Infectious diseases | no | 0 | 581,304 | 53.06 |
| | yes | 1 | 514,309 | 46.94 |
| Neoplasia | no | 0 | 961,674 | 87.78 |
| | yes | 1 | 133,939 | 12.23 |

| | | | | |
|---|---|---|---|---|
| Blood disease | no | 0 | 659,726 | 60.22 |
| | yes | 1 | 435,887 | 39.78 |
| Nerve disease | no | 0 | 698,471 | 63.75 |
| | yes | 1 | 397,142 | 36.25 |
| Diseases of the respiratory system | no | 0 | 878,327 | 80.17 |
| | yes | 1 | 217,286 | 19.83 |
| Muscular diseases | no | 0 | 110,723 | 10.11 |
| | yes | 1 | 984,890 | 89.89 |
| Diseases of the urinary tract | no | 0 | 1,034,044 | 94.38 |
| | yes | 1 | 61,569 | 5.62 |
| Diseases of the digestive tract | no | 0 | 1,035,118 | 94.48 |
| | yes | 1 | 60,495 | 5.52 |
| Endocrine diseases | no | 0 | 871,693 | 79.56 |
| | yes | 1 | 223,920 | 20.44 |
| diseases of the circulatory system | no | 0 | 683,534 | 62.39 |
| | yes | 1 | 412,079 | 37.61 |
| Metabolism diseases | no | 0 | 431,410 | 39.38 |
| | yes | 1 | 664,203 | 60.62 |
| Cardiovascular diseases | no | 0 | 299,798 | 27.36 |
| | yes | 1 | 795,815 | 72.64 |
| Drugs for dermatological problems | no | 0 | 1,068,253 | 97.50 |
| | yes | 1 | 27,360 | 2.50 |
| Genital diseases | no | 0 | 946,439 | 86.38 |
| | yes | 1 | 149,174 | 13.62 |
| Hormonal diseases | no | 0 | 839,661 | 76.64 |
| | yes | 1 | 255,952 | 23.36 |
| Problems with the sensory parts | no | 0 | 1,019,402 | 93.04 |
| | yes | 1 | 76,211 | 6.96 |

# B Chi-square Test Results between Samples

| Variable Name | Code | Sample-1 (10%) | | Sample-2 (90%) | | $X^2$ | DF | CV | *P*-values |
|---|---|---|---|---|---|---|---|---|---|
| | | Count | % | N | % | | | | |
| Age | 0 | 27,897 | 26 | 234,694 | 27 | 4.36 | 5 | 11.07 | .50 |
| | 1 | 23,512 | 22 | 195,078 | 22 | | | | |
| | 2 | 22,912 | 22 | 191,267 | 22 | | | | |
| | 3 | 16,560 | 16 | 136,996 | 16 | | | | |
| | 4 | 10,113 | 10 | 84,746 | 10 | | | | |
| | 5 | 4,968 | 5 | 40,908 | 5 | | | | |
| Citizenship | 0 | 104,088 | 98 | 867,665 | 98 | 1.07 | 1 | 3.84 | .30 |
| | 1 | 1,874 | 2 | 16,024 | 2 | | | | |
| Number of urgent hospitalization | 0 | 100,886 | 95 | 840,603 | 95 | 1.90 | 2 | 5.99 | .39 |
| | 1 | 4,931 | 5 | 41,785 | 5 | | | | |
| | 2 | 145 | 0 | 1,301 | 0 | | | | |
| Number of nontraumatic hospitalizations | 0 | 97,230 | 92 | 809,990 | 92 | 2.65 | 2 | 5.99 | .27 |
| | 1 | 8,149 | 8 | 69,023 | 8 | | | | |
| | 2 | 583 | 1 | 4,676 | 1 | | | | |
| Number of total hospitalizations | 0 | 96,423 | 91 | 803,210 | 91 | 1.86 | 2 | 5.99 | .39 |
| | 1 | 8,840 | 8 | 74,757 | 8 | | | | |
| | 2 | 699 | 1 | 5,722 | 1 | | | | |
| Charlson index | 0 | 100,247 | 95 | 836,048 | 95 | 0.69 | 3 | 7.81 | .88 |
| | 1 | 4,836 | 5 | 40,120 | 5 | | | | |
| | 2 | 627 | 1 | 5,392 | 1 | | | | |
| | 3 | 252 | 0 | 2,129 | 0 | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Home-based care | 0 | 104,868 | 99 | 874,682 | 99 | 0.16 | 1 | 3.84 | .69 |
| | 1 | 1,094 | 1 | 9,007 | 1 | | | | |
| Income | 0 | 24,990 | 24 | 208,621 | 24 | 0.03 | 1 | 3.84 | .86 |
| | 1 | 80,972 | 76 | 675,068 | 76 | | | | |
| Invalidity | 0 | 94,608 | 89 | 789,101 | 89 | 0.01 | 1 | 3.84 | .91 |
| | 1 | 11,354 | 11 | 94,588 | 11 | | | | |
| Poly prescriptions (number of drugs prescribed) | 0 | 56,475 | 53 | 470,901 | 53 | 2.50 | 2 | 5.99 | .29 |
| | 1 | 35,262 | 33 | 292,723 | 33 | | | | |
| | 2 | 14,225 | 13 | 120,065 | 14 | | | | |
| Number of different types of drugs prescribed ( First three digits of ATC code) | 0 | 8,573 | 8 | 70,642 | 8 | 1.25 | 3 | 7.81 | .74 |
| | 1 | 56,665 | 53 | 472,961 | 54 | | | | |
| | 2 | 40,254 | 38 | 336,201 | 38 | | | | |
| | 3 | 470 | 0 | 3,885 | 0 | | | | |
| Access to ED with White code | 0 | 102,833 | 97 | 857,402 | 97 | 0.16 | 1 | 3.84 | .69 |
| | 1 | 3,129 | 3 | 26,287 | 3 | | | | |
| Access to ED with Green code | 0 | 86,958 | 82 | 724,742 | 82 | 0.34 | 2 | 5.99 | .84 |
| | 1 | 17,399 | 16 | 145,384 | 16 | | | | |
| | 2 | 1,605 | 2 | 13,563 | 2 | | | | |
| Access to ED with Yellow code | 0 | 101,762 | 96 | 847,565 | 96 | 3.82 | 2 | 5.99 | .15 |
| | 1 | 4,122 | 4 | 35,477 | 4 | | | | |
| | 2 | 78 | 0 | 647 | 0 | | | | |
| Access to ED with Red code | 0 | 105,740 | 100 | 881,738 | 100 | 0.55 | 1 | 3.84 | .46 |
| | 1 | 222 | 0 | 1,951 | 0 | | | | |
| Housing condition | 1 | 86,678 | 82 | 722,002 | | 0.55 | 2 | 5.99 | .76 |
| | 2 | 14,853 | 14 | 123,963 | 14 | | | | |
| | 3 | 4,431 | 4 | 37,724 | 4 | | | | |
| Marital status | 1 | 7,730 | 7 | 65,119 | 7 | 1.10 | 3 | 7.81 | .78 |
| | 2 | 68,656 | 65 | 572,836 | 65 | | | | |
| | 3 | 23,153 | 22 | 192,280 | 22 | | | | |
| | 4 | 6,423 | 6 | 53,454 | 6 | | | | |
| Level of education | 1 | 18,736 | 18 | 157,134 | 18 | 0.75 | 2 | 5.99 | .69 |
| | 3 | 32,145 | 30 | 268,151 | 30 | | | | |
| | 4 | 55,081 | 52 | 458,404 | 52 | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Work status | 1 | 9,680 | 9 | 80,955 | 9 | 0.29 | 3 | 7.81 | .96 |
| | 3 | 10,155 | 10 | 84,838 | 10 | | | | |
| | 5 | 81,694 | 77 | 680,710 | 77 | | | | |
| | 7 | 4,433 | 4 | 37,186 | 4 | | | | |
| Home living status | 1 | 98,811 | 93 | 823,362 | 93 | 0.90 | 1 | 3.84 | .34 |
| | 2 | 7,151 | 7 | 60,327 | 7 | | | | |
| Type of family | 1 | 36,780 | 35 | 305,745 | 35 | 1.66 | 3 | 7.81 | .65 |
| | 2 | 6,147 | 6 | 50,654 | 6 | | | | |
| | 3 | 33,482 | 32 | 280,202 | 32 | | | | |
| | 4 | 29,553 | 28 | 247,088 | 28 | | | | |
| Disability | 0 | 93,636 | 88 | 781,148 | 88 | 0.08 | 1 | 3.84 | .78 |
| | 1 | 12,326 | 12 | 102,541 | 12 | | | | |
| Femur Fracture | 0 | 101,331 | 96 | 843,904 | 96 | 3.83 | 1 | 3.84 | .05 |
| | 1 | 4,631 | 4 | 39,785 | 5 | | | | |
| Depression | 0 | 91,148 | 86 | 759,378 | 86 | 0.59 | 1 | 3.84 | .44 |
| | 1 | 14,814 | 14 | 124,311 | 14 | | | | |
| Diabetes | 0 | 89,437 | 84 | 744,236 | 84 | 2.45 | 1 | 3.84 | .12 |
| | 1 | 16,525 | 16 | 139,453 | 16 | | | | |
| Arthropathy | 0 | 101,320 | 96 | 845,264 | 96 | 0.24 | 1 | 3.84 | .62 |
| | 1 | 4,642 | 4 | 38,425 | 4 | | | | |
| Parkinson's disease | 0 | 104,233 | 98 | 869,700 | 98 | 1.44 | 1 | 3.84 | .23 |
| | 1 | 1,729 | 2 | 13,989 | 2 | | | | |
| Epilepsy | 0 | 100,577 | 95 | 838,499 | 95 | 0.20 | 1 | 3.84 | .66 |
| | 1 | 5,385 | 5 | 45,190 | 5 | | | | |
| Anemia | 0 | 99,657 | 94 | 830,663 | 94 | 0.42 | 1 | 3.84 | .51 |
| | 1 | 6,305 | 6 | 53,026 | 6 | | | | |
| Hypertensive | 0 | 76,533 | 72 | 637,821 | 72 | 0.12 | 1 | 3.84 | .73 |
| | 1 | 29,429 | 28 | 245,868 | 28 | | | | |
| Glaucoma | 0 | 98,385 | 93 | 820,530 | 93 | 0.00 | 1 | 3.84 | .97 |
| | 1 | 7,577 | 7 | 63,159 | 7 | | | | |
| Mental disease | 0 | 103,095 | 97 | 860,158 | 97 | 0.67 | 1 | 3.84 | .41 |
| | 1 | 2,867 | 3 | 23,531 | 3 | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Cancer | 0 | 94,429 | 89 | 787,377 | 89 | 0.02 | 1 | 3.84 | .88 |
| | 1 | 11,533 | 11 | 96,312 | 11 | | | | |
| Thyroid | 0 | 96,048 | 91 | 801,056 | 91 | 0.00 | 1 | 3.84 | .96 |
| | 1 | 9,914 | 9 | 82,633 | 9 | | | | |
| Dementia | 0 | 104,948 | 99 | 875,171 | 99 | 0.05 | 1 | 3.84 | .83 |
| | 1 | 1,014 | 1 | 8,518 | 1 | | | | |
| Coronary artery disease | 0 | 101,713 | 96 | 848,154 | 96 | 0.03 | 1 | 3.84 | .86 |
| | 1 | 4,249 | 4 | 35,535 | 4 | | | | |
| Congestive heart failure | 0 | 105,099 | 99 | 876,572 | 99 | 0.10 | 1 | 3.84 | .76 |
| | 1 | 863 | 1 | 7,117 | 1 | | | | |
| Stroke | 0 | 103,549 | 98 | 863,804 | 98 | 0.31 | 1 | 3.84 | .58 |
| | 1 | 2,413 | 2 | 19,885 | 2 | | | | |
| kidney failure | 0 | 105,048 | 99 | 875,966 | 99 | 0.14 | 1 | 3.84 | .71 |
| | 1 | 914 | 1 | 7,723 | 1 | | | | |
| Hypercholesterolemia | 0 | 103,254 | 97 | 861,238 | 97 | 0.09 | 1 | 3.84 | .77 |
| | 1 | 2,708 | 3 | 22,451 | 3 | | | | |
| Atrial fibrillation | 0 | 101,702 | 96 | 848,265 | 96 | 0.03 | 1 | 3.84 | .86 |
| | 1 | 4,260 | 4 | 35,424 | 4 | | | | |
| Neck fracture | 0 | 105,117 | 99 | 876,678 | 99 | 0.02 | 1 | 3.84 | .89 |
| | 1 | 845 | 1 | 7,011 | 1 | | | | |
| Infectious diseases | 0 | 57,484 | 54 | 478,823 | 54 | 0.16 | 1 | 3.84 | .69 |
| | 1 | 48,478 | 46 | 404,866 | 46 | | | | |
| Neoplasia | 0 | 93,736 | 88 | 781,662 | 88 | 0.01 | 1 | 3.84 | .94 |
| | 1 | 12,226 | 12 | 102,027 | 12 | | | | |
| Blood disease | 0 | 66,208 | 62 | 553,085 | 63 | 0.45 | 1 | 3.84 | .50 |
| | 1 | 39,754 | 38 | 330,604 | 37 | | | | |
| Nerve disease | 0 | 69,698 | 66 | 580,218 | 66 | 0.58 | 1 | 3.84 | .45 |
| | 1 | 36,264 | 34 | 303,471 | 34 | | | | |
| Diseases of the respiratory system | 0 | 86,452 | 82 | 720,816 | 82 | 0.02 | 1 | 3.84 | .88 |
| | 1 | 19,510 | 18 | 162,873 | 18 | | | | |
| Muscular diseases | 0 | 9,649 | 9 | 80,108 | 9 | 0.19 | 1 | 3.84 | .66 |
| | 1 | 96,313 | 91 | 803,581 | 91 | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Diseases of the urinary tract | 0 | 100,843 | 95 | 841,242 | 95 | 0.16 | 1 | 3.84 | 0.69 |
| | 1 | 5,119 | 5 | 42,447 | 5 | | | | |
| Diseases of the digestive tract | 0 | 100,932 | 95 | 840,559 | 95 | 3.65 | 1 | 3.84 | .06 |
| | 1 | 5,030 | 5 | 43,130 | 5 | | | | |
| Endocrine diseases | 0 | 85,300 | 81 | 710,744 | 80 | 0.31 | 1 | 3.84 | .58 |
| | 1 | 20,662 | 20 | 172,945 | 20 | | | | |
| Diseases of the circulatory system | 0 | 68,293 | 64 | 569,475 | 64 | 0.00 | 1 | 3.84 | .96 |
| | 1 | 37,669 | 36 | 314,214 | 36 | | | | |
| Metabolism diseases | 0 | 43,419 | 41 | 360,291 | 41 | 1.64 | 1 | 3.84 | .20 |
| | 1 | 62,543 | 59 | 523,398 | 59 | | | | |
| Cardiovascular diseases | 0 | 30,111 | 28 | 250,030 | 28 | 0.70 | 1 | 3.84 | .40 |
| | 1 | 75,851 | 72 | 633,659 | 72 | | | | |
| Drugs for dermatological problems | 0 | 103,420 | 98 | 862,639 | 98 | 0.12 | 1 | 3.84 | .73 |
| | 1 | 2,542 | 2 | 21,050 | 2 | | | | |
| Genital diseases | 0 | 91,741 | 87 | 765,505 | 87 | 0.18 | 1 | 3.84 | .67 |
| | 1 | 14,221 | 13 | 118,184 | 13 | | | | |
| Hormonal diseases | 0 | 82,094 | 77 | 683,505 | 77 | 0.89 | 1 | 3.84 | .35 |
| | 1 | 23,868 | 23 | 200,184 | 23 | | | | |
| Problems with the sensory parts | 0 | 98,687 | 93 | 822,827 | 93 | 0.07 | 1 | 3.84 | .79 |
| | 1 | 7,275 | 7 | 60,862 | 7 | | | | |

# C Parameters Settings for Machine learning Algorithms

| Hyperparameters settings for each ML model in each of the six problems | | | | | |
|---|---|---|---|---|---|
| **Problem** | SVM | ANN | RF | DT | LR |
| Mortality | 'C': 100, 'gamma': 0.001, 'kernel': 'rbf' | 'activation': 'relu', 'alpha': 0.05, 'hidden_layer_sizes': (30, 30, 30), 'learning_rate': 'constant', 'solver': 'sgd' | 'max_depth': 90, 'max_features': 3, 'min_samples_leaf': 5, 'min_samples_split': 12, 'n_estimators': 100 | 'criterion': 'entropy', 'max_depth': 12, 'max_features': 5, 'min_samples_split': 2 | 'C': 10.0, 'penalty': 'l2' |
| Fracture | 'C': 1, 'gamma': 0.01, 'kernel': 'rbf' | 'activation': 'relu', 'alpha': 0.05, 'hidden_layer_sizes': (50, 100, 50), 'learning_rate': 'adaptive', 'solver': 'sgd' | 'max_depth': 80, 'max_features': 8, 'min_samples_leaf': 5, 'min_samples_split': 12, 'n_estimators': 200 | 'criterion': 'entropy', 'max_depth': 6, 'max_features': 30, 'min_samples_split': 5 | 'C': 1000.0, 'penalty': 'l2' |

| | | | | | |
|---|---|---|---|---|---|
| Disability | 'C': 1, 'gamma': 0.1, 'kernel': 'rbf' | 'activation': 'tanh', 'alpha': 0.05, 'learning_rate': 'constant', 'solver':'sgd,hidden_layer_sizes':(10,10,10) | 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 8, 'n_estimators': 100, 'max_depth':80 | 'max_depth': 100, 'max_features': 3, 'min_samples_split': 12, 'n_estimators': 100 | 'C': 100.0, 'penalty': 'l2' |
| Urgent Hosp. | 'C': 100, 'gamma': 0.001, 'kernel': 'rbf' | 'activation': 'relu', 'alpha': 0.0001, 'learning_rate': 'constant', 'solver':'adam, hidden_layer_sizes':(100,) | 'max_features': 3, 'min_samples_leaf': 5, 'min_samples_split': 12, 'n_estimators': 100, 'max_depth':90 | 'criterion': 'entropy', 'max_depth': 6, 'max_features': 30, 'min_samples_split': 2 | 'C': 1.0, 'penalty': 'l2' |

| | | | | | |
|---|---|---|---|---|---|
| Preventable Hosp. | 'C': 10, 'gamma': 0.001, 'kernel': 'rbf' | 'activation': 'tanh', 'alpha': 0.05, 'hidden_layer_sizes': (100,), 'learning_rate': 'constant', 'solver': 'sgd' | 'max_features': 3, 'min_samples_leaf': 4, 'min_samples_split': 12, 'n_estimators': 300, 'max_depth':90 | 'criterion': 'gini', 'max_depth': 6, 'max_features': 10, 'min_samples_split': 2 | 'C': 10.0, 'penalty': 'l2' |
| Access to emergency department with red code | 'C': 1, 'gamma': 0.1, 'kernel': 'rbf' | 'activation': 'relu', 'alpha': 0.0001, 'learning_rate': 'adaptive', 'solver': 'sgd' | 'max_depth': 80, 'max_features': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 300 | 'criterion': 'entropy', 'max_depth': 6, 'max_features': 40, 'min_samples_split': 2 | 'C': 0.1, 'penalty': 'l2' |

# D Feature Selection

D.1: The most important variables in disability and urgent hospitalization problems

| **Disability problem** | | | **Urgent hospitalization problem** | | |
|---|---|---|---|---|---|
| **Variable** | **Rank** | ***P*-Value** | **Variable** | **Rank** | ***P*-Value** |
| Age | 1 | *P*<.001 | Age | 1 | *P*<.001 |
| Charlson index | 2 | *P*<.001 | Mental Disease | 2 | *P*<.001 |
| # total hospitalizations | 3 | *P*<.001 | Poly prescriptions | 3 | *P*<.001 |
| # urgent hospitalization | 4 | *P*<.001 | diseases of the respiratory system | 4 | *P*<.001 |
| Poly prescriptions | 5 | *P*<.001 | Citizenship | 5 | *P*<.001 |
| # non-traumatic | 6 | *P*<.001 | White code | 6 | *P*<.001 |
| Green code | 7 | *P*<.001 | arthropathy | 7 | *P*<.001 |
| Nerve disease | 8 | *P*<.001 | diseases of the circulatory system | 8 | *P*<.001 |
| Disability | 10 | *P*<.001 | Glaucoma | 10 | *P*<.001 |
| blood disease | 11 | *P*<.001 | Femur fracture | 11 | *P*<.001 |
| Yellow code | 12 | *P*<.001 | Heart disease | 12 | *P*<.001 |

| | | | | | |
|---|---|---|---|---|---|
| Depression | 13 | *P*<.001 | Nerve disease | 13 | *P*<.001 |
| diseases of the circulatory system | 14 | *P*<.001 | Neoplasm | 14 | *P*<.001 |
| Dementia | 15 | *P*<.001 | Disability | 15 | *P*<.001 |
| Anaemia | 16 | *P*<.001 | Drugs for dermatological problems | 16 | *P*<.001 |
| mental disease | 17 | *P*<.001 | metabolism diseases | 17 | *P*<.001 |
| diseases of the urinary tract | 18 | *P*<.001 | genital diseases | 18 | *P*<.001 |
| Parkinson disease | 19 | *P*<.001 | hormonal diseases | 19 | *P*<.001 |

**D.2:** The most important variables in preventable hospitalization and emergency admission

| **Preventable hospitalization** | | | **Emergency admission with red code** | | |
|---|---|---|---|---|---|
| **Variable** | **Rank** | **P-Value** | **Variable** | **Rank** | **P-Value** |
| Age | 1 | *P*<.001 | Age | 1 | *P*<.001 |
| Mental disease | 2 | *P*<.001 | Charlson index | 2 | *P*<.001 |
| Poly prescriptions | 3 | *P*<.001 | # urgent hospitalization | 3 | *P*<.001 |
| diseases of the respiratory system | 4 | *P*<.001 | # total hospitalization | 4 | *P*<.001 |
| White code | 5 | *P*<.001 | Poly prescriptions | 5 | *P*<.001 |
| Citizenship | 6 | *P*<.001 | # non-traumatic | 6 | *P*<.001 |

| | | | | | |
|---|---|---|---|---|---|
| arthropathy | 7 | *P*<.001 | Yellow code | 7 | *P*<.001 |
| diseases of the circulatory system | 8 | *P*<.001 | Invalidity | 8 | *P*<.001 |
| Glaucoma | 10 | *P*<.001 | Disability | 10 | *P*<.001 |
| Heat disease | 12 | *P*<.001 | diseases of the respiratory system | 12 | *P*<.001 |
| Femur fracture | 13 | *P*<.001 | Blood disease | 13 | *P*<.001 |
| Nerve disease | 14 | *P*<.001 | diseases of the circulatory system | 14 | *P*<.001 |
| Neoplasm | 15 | *P*<.001 | Green code | 15 | *P*<.001 |
| metabolism diseases | 16 | *P*<.001 | diseases of the urinary tract | 16 | *P*<.001 |
| Drugs for dermatological problems | 17 | *P*<.001 | Anaemia | 17 | *P*<.001 |
| drugs for the sensory parts | 18 | *P*<.001 | Congestive heart failure | 18 | *P*<.001 |

# E Algorithm for ML-TLSMOTE

1. Start with an input dataset *D* // Pre-process minority samples using SMOTE
2. L1 ←LabelsInDataset(D) // get the set of all labels in D
3. MeanIR ←GetMeanIR(D) // get MeanIR of labels in D
4. **for each** label **in** L1
    a. IRLbl ←getIRLbl (label) // imbalance ratio of each label
    b. **If** (IRLbl) > MeanIR) **then**
        i. minBags(label) ← GetInstances(label) // bags of minority label samples
    **c. end if**
5. For each minBag in minBags
    a. **for each** sample in minBag and *lb* in L1
        i. T ← instances that are associated and non-associated with *lb*, but only non-associated with the other combinations of labels
        ii. A ← D \ T , where D \ T is the set difference // keep all the remaining instances of $x$ in A , s.t , $x \in$ D and $x \notin$ T
        iii. Set the imbalance percentage
        iv. **for each** point $p$ in T,
            1. Compute its k nearest neighbours in T
            2. Randomly choose $r \leq k$ of the neighbours (with replacement).
            3. Generate synthetic instance along the lines joining $p$ and each of the $r$ selected neighbours.
            4. Add the generated synthetic instance to the feature vector and labelset of $p$
            5. S ← synthetic samples found in step (4) U T
        v. W ←S U A // W is the union of S and A

6. Start from Dataset W // Pre-process majority samples using Tomek links
7. L2 ←LabelsInDataset(W)
8. DT ←samples that are associated with at least one label in L2
9. L3 ←LabelsInDataset(DT)
10. MeanIR ←GetMeanIR(DT)
11. **for each** label **in** L3 **do**
    a. IRLbl ←getIRLbl (label)
    b. **If** (IRLbl < MeanIR) **then**
        i. majBag ← GetAllInstances(label)
    **c. end if**
12. **end for**
13. TL ← empty list of instances
14. A ←instances of negative class in majBag) // class 0
15. B ←instances of positive class in majBag) // class 1
16. **for each** sample **in** majBag
    a. **if** sample ∈ A
        i. i ←sample
    **b. else**
        i. j ←sample // i.e j∈ B
    c. d(i,j) ←Distance (i, j)
    **d. end if**
        i. **for any** sample E **in** (A U B)
            1. **if** (d(i,j) < d(i,E) or d(i,j) <d(j,E))
                a. TL ← (i,j) // mark a pair (i,j) is a Tomek link
                **b.** F←W-TL // remove TL from W
            **2. end if**
        **ii. end for**
17. **end for**
18. **return** F //a pre-processed dataset

# Bibliography


[1] United Nations, World Population Ageing 2017 - Highlights, 2017.

[2] Eurostat, People in the EU: who are we and how do we live? , 2015. https://ec.europa.eu/eurostat/web/products-statistical-books/-/KS-04-15-567.

[3] World Health Organization, Are you ready? What you need to know about ageing, (2012) 1–6. http://www.who.int/world-health-day/2012/toolkit/background/en/index3.html.

[4] World Health Organisation, Why Population Ageing Matters to Health: A Global Perspective, 2011.

[5] K. Rockwood, S.E. Howlett, Fifteen years of progress in understanding frailty and health in aging, BMC Medicine. (2018). doi:https://doi.org/10.1186/s12916-018-1223-3.

[6] D. Stow, F.E. Matthews, S. Barclay, S. Iliffe, A. Clegg, S. De Biase, L. Robinson, B. Hanratty, Evaluating frailty scores to predict mortality in older adults using data from population based electronic health records: Case control study, Age and Ageing. (2018).doi:https://doi.org/10.1093/ageing/afy022.

[7] G. Kojima, A. Liljas, S. Iliffe, Frailty syndrome: implications and challenges for health care policy, Risk Management and Healthcare Policy. Volume 12 (2019) 23–30.doi:https://doi.org/10.2147/RMHP.S168750.

[8] T.E. Strandberg, K.H. Pitkälä, R.S. Tilvis, Frailty in older people, European Geriatric Medicine. (2011). doi:https://doi.org/10.1016/j.eurger.2011.08.003.

[9] R.J.J. Gobbens, K.G. Luijkx, M.T. Wijnen-Sponselee, J.M.G.A. Schols, Towards an integral conceptual model of frailty, Journal of Nutrition, Health and Aging. (2010). doi:https://doi.org/10.1007/s12603-010-0045-6.

[10] L.P. Fried, L. Ferrucci, J. Darer, J.D. Williamson, G. Anderson, Untangling the Concepts of Disability, Frailty, and Comorbidity: Implications

for Improved Targeting and Care, The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 59 (2004) M255–M263. doi:https://doi.org/10.1093/gerona/59.3.M255.

[11] M. Markle-Reid, G. Browne, Conceptualizations of frailty in relation to older adults, Journal of Advanced Nursing. (2003).doi:https://doi.org/10.1046/j.1365-2648.2003.02767.x.

[12] J.P.J. Slaets, Vulnerability in the Elderly: Frailty, Medical Clinics of North America. (2006). doi:https://doi.org/10.1016/j.mcna.2006.05.008.

[13] A. Clegg, J. Young, E.S. Iliff, M.O. Rikkert, K. Rockwood, Erratum: Frailty in elderly people (The Lancet (2013) 381 (752-62)), The Lancet. (2013).doi:https://doi.org/10.1016/S0140-6736(13)62138-8.

[14] T.B.L. Kirkwood, Understanding the odd science of aging, Cell. (2005).doi:https://doi.org/10.1016/j.cell.2005.01.027.

[15] N. Fairhall, C. Langron, C. Sherrington, S.R. Lord, S.E. Kurrle, K. Lockwood, N. Monaghan, C. Aggar, L. Gill, I.D. Cameron, Treating frailty-a practical guide, BMC Medicine. (2011).doi:https://doi.org/10.1186/1741-7015-9-83.

[16] I.D. Cameron, N. Fairhall, C. Langron, K. Lockwood, N. Monaghan, C. Aggar, C. Sherrington, S.R. Lord, S.E. Kurrle, A multifactorial interdisciplinary intervention reduces frailty in older people: Randomized trial, BMC Medicine. (2013). doi:https://doi.org/10.1186/1741-7015-11-65.

[17] T.P. Ng, L. Feng, M.S.Z. Nyunt, L. Feng, M. Niti, B.Y. Tan, G. Chan, S.A. Khoo, S.M. Chan, P. Yap, K.B. Yap, Nutritional, Physical, Cognitive, and Combination Interventions and Frailty Reversal among Older Adults: A Randomized Controlled Trial, American Journal of Medicine. (2015).doi:https://doi.org/10.1016/j.amjmed.2015.06.017.

[18] E.P. Cherniack, H.J. Flores, B.R. Troen, Emerging therapies to treat frailty syndrome in the elderly, Alternative Medicine Review. (2007).

[19] K.T. Galvin, L. Todres, Kinds of well-being: A conceptual framework that provides direction for caring, International Journal of Qualitative Studies on Health and Well-Being. (2011). doi:https://doi.org/10.3402/qhw.v6i4.10362.

[20] L.P. Fried, C.M. Tangen, J. Walston, A.B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W.J. Kop, G. Burke, M.A. McBurnie, Frailty in Older Adults: Evidence for a Phenotype, The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 56 (2001) M146–M157. doi:https://doi.org/10.1093/gerona/56.3.M146.

[21] K. Rockwood, X. Song, C. MacKnight, H. Bergman, D.B. Hogan, I. McDowell, A. Mitnitski, A global clinical measure of fitness and frailty in elderly people, CMAJ. (2005).doi:https://doi.org/10.1503/cmaj.050051.

[22] Y.N. Panhwar, F. Naghdy, G. Naghdy, D. Stirling, J. Potter, Assessment of frailty: a survey of quantitative and clinical methods, BMC Biomedical Engineering. 1 (2019) 7.doi:https://doi.org/10.1186/s42490-019-0007-y.

[23] J. Soong, A.J. Poots, S. Scott, K. Donald, T. Woodcock, D. Lovett, D. Bell, Quantifying the prevalence of frailty in English hospitals, BMJ Open. (2015).doi:https://doi.org/10.1136/bmjopen-2015-008456.

[24] V. Fillion, M.J. Sirois, P. Gamache, J.R. Guertin, S.N. Morin, S. Jean, Frailty and health services use among Quebec seniors with non-hip fractures: A population-based study using adminsitrative databases, BMC Health Services Research. (2019). doi:https://doi.org/10.1186/s12913-019-3865-z.

[25] P. Grimm, Social Desirability Bias, in: Wiley International Encyclopedia of Marketing, 2010. doi:https://doi.org/10.1002/9781444316568.wiem02057.

[26] R. Pirracchio, M.J. Cohen, I. Malenica, J. Cohen, A. Chambaz, M. Cannesson, C. Lee, M. Resche-Rigon, A. Hubbard, Big data and targeted machine learning in action to assist medical decision in the ICU, Anaesthesia Critical Care and Pain Medicine. (2019). https://doi.org/10.1016/j.accpm.2018.09.008.

[27] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: Past, present and future, Stroke and Vascular Neurology. (2017). doi:https://doi.org/10.1136/svn-2017-000101.

[28] Song X, Mitnitski A, MacKnight C, et al. Assessment of individual risk of death using selfreport data: an artificial neural network compared with a frailty index. J Am Geriatr Soc 2004;52:1180–4. doi:https://doi:10.1111/j.1532-5415.2004.52319.x

[29] Shao Y, Mohanty AF, Ahmed A, et al. Identification and Use of Frailty Indicators from Text to Examine Associations with Clinical Outcomes Among Patients with Heart Failure.AMIA Annu Symp Proc 2016:1110–8

[30] Chen T, Dredze M, Weiner JP, et al. Extraction of Geriatric Syndromes

From Electronic Health Record Clinical Notes: Assessment of Statistical Natural Language Processing Methods. JMIR Med Informatics 2019;7:e13039. doi:https://doi:10.2196/13039

[31] Anzaldi LJ, Davison A, Boyd CM, et al. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records, a restrospective cohort study. BMC Geriatr 2017;17:248. doi:https://doi:10.1186/s12877-017-0645-7

[32] Kuo K-M, Talley PC, Kuzuya M, et al. Development of a clinical support system for identifying social frailty. Int J Med Inform 2019;132:103979. doi:https://doi:10.1016/j.ijmedinf.2019.103979

[33] Fabrice Mowbray Manaf Zargoush et al. Predicting hospital admission for older emergency department patients: Insights from machine learning, International Journal of Medical Informatics (2020). doi:https://doi.org/10.1016/j.ijmedinf.2020.104163

[34] A. Clegg, J. Young, Frailty and Organization of Health and Social Care, Interdisciplinary Topics in Gerontology and Geriatrics. (2015). doi:https://doi.org/10.1159/000381233.

[35] J. LYNN, B.M. STRAUBE, K.M. BELL, S.F. JENCKS, R.T. KAMBIC, Using Population Segmentation to Provide Better Health Care for All: The "Bridges to Health" Model, The Milbank Quarterly. 85 (2007) 185–208. doi:https://doi.org/10.1111/j.1468-0009.2007.00483.x.

[36] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, 2012. doi:https://doi.org/10.1016/C2009-0-61819-5.

[37] A.M. Mudge, R.E. Hubbard, Frailty: Mind the gap, Age and Ageing. (2018).doi:https://doi.org/10.1093/ageing/afx193.

[38] K. Rockwood, R.A. Fox, P. Stolee, D. Robertson, B.L. Beattie, Frailty in elderly people: An evolving concept, CMAJ. (1994).

[39] ONU, World population, ageing, Suggested Citation: United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Ageing. (2015). doi:https://doi.org/ST/ESA/SER.A/390.

[40] E. Joosten, M. Demuynck, E. Detroyer, K. Milisen, Prevalence of frailty and its ability to predict in hospital delirium, falls, and 6-month mortality in hospitalized older patients, BMC Geriatrics. (2014). doi:https://doi.org/10.1186/1471-2318-14-1.

[41] Y. Buist, M. Rijken, L. Lemmens, C. Baan, S. de Bruin, Collaborating on early detection of frailty; A multifaceted challenge, International Journal of Integrated Care. (2019).

doi:https://doi.org/10.5334/ijic.4176.

[42] World Health Organisation, WHO Clinical Consortium on Healthy Ageing. Report of consortium meeting 1-2 December 2016 in Geneva, Switzerland, World Health Organisation: Clinical Consortium on Healthy Ageing. (2017).

[43] A.B. Mitnitski, A.J. Mogilner, K. Rockwood, Accumulation of deficits as a proxy measure of aging., TheScientificWorldJournal. (2001). doi:https://doi.org/10.1100/tsw.2001.58.

[44] G.A. Aguayo, M.T. Vaillant, A.F. Donneau, A. Schritz, S. Stranges, L. Malisoux, A. Chioti, M. Guillaume, M. Muller, D.R. Witte, Comparative analysis of the association between 35 frailty scores and cardiovascular events, cancer, and total mortality in an elderly general population in England: An observational study, PLoS Medicine. (2018). doi:https://doi.org/10.1371/journal.pmed.1002543.

[45] O. Theou, M.R.H. Rockwood, A. Mitnitski, K. Rockwood, Disability and co-morbidity in relation to frailty: How much do they overlap?, Archives of Gerontology and Geriatrics. (2012). doi:https://doi.org/10.1016/j.archger.2012.03.001.

[46] S. Katz, T.D. Downs, H.R. Cash, R.C. Grotz, Progress in development of the index of ADL., The Gerontologist. (1970). doi:https://doi.org/10.1093/geront/10.1_Part_1.20.

[47] M.P. Lawton, E.M. Brody, Assessment of older people: Self-maintaining and instrumental activities of daily living, Gerontologist. (1969).doi:https://doi.org/10.1093/geront/9.3_Part_1.179.

[48] C.T. Cigolle, M.B. Ofstedal, Z. Tian, C.S. Blaum, Comparing models of frailty: The health and retirement study, Journal of the American Geriatrics Society. (2009).doi:https://doi.org/10.1111/j.1532-5415.2009.02225.x.

[49] A.B. Mitnitski, J.E. Graham, A.J. Mogilner, K. Rockwood, Frailty, fitness and late-life mortality in relation to chronological and biological age, BMC Geriatrics. (2002). doi:https://doi.org/10.1186/1471-2318-2-1.

[50] P.K. Myint, A.A. Welch, Healthier ageing, BMJ (Online). (2012). doi:https://doi.org/10.1136/bmj.e1214.

[51] L.A. Lipsitz, Dynamic models for the study of frailty, Mechanisms of Ageing and Development. (2008). doi:https://doi.org/10.1016/j.mad.2008.09.012.

[52] H. Schuurmans, N. Steverink, S. Lindenberg, N. Frieswijk, J.P.J.

Slaets, Old or Frail: What Tells Us More?, The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. (2004). doi:https://doi.org/10.1093/gerona/59.9.m962.

[53] M.J. Haapanen, M.M., M.K. Salonen, E. Kajantie, M. Simonen, P. Pohjolainen, J.G. Eriksson, M.B. Von Bonsdorff, Early life determinants of frailty in old age: The Helsinki birth cohort study, Age and Ageing. (2018). doi:https://doi.org/10.1093/ageing/afy052.

[54] K. Rockwood, D.B. Hogan, C. MacKnight, Conceptualisation and measurement of frailty in elderly people, Drugs and Aging. (2000). doi:https://doi.org/10.2165/00002512-200017040-00005.

[55] F. Buckinx, Y. Rolland, J.Y. Reginster, C. Ricour, J. Petermans, O. Bruyère, Burden of frailty in the elderly population: Perspectives for a public health challenge, Archives of Public Health. (2015).doi:https://doi.org/10.1186/s13690-015-0068-x.

[56] M.K. Andrew, S. Dupuis-Blanchard, C. Maxwell, A. Giguere, J. Keefe, K. Rockwood, P. St John, Social and societal implications of frailty, including impact on Canadian healthcare systems, The Journal of Frailty and Aging. (2018). doi:https://doi.org/10.14283/jfa.2018.30.

[57] G. Kojima, Frailty Defined by FRAIL Scale as a Predictor of Mortality: A Systematic Review and Meta-analysis, Journal of the American Medical Directors Association. (2018). doi:https://doi.org/10.1016/j.jamda.2018.04.006.

[58] Gotaro Kojima, Increased healthcare costs associated with frailty among community-dwelling older people: A systematic review and meta-analysis, Archives of Gerontology and Geriatrics. 84 (2019) 103898. https://doi.org/10.1016/j.archger.2019.06.003.

[59] P.D. De Jonge, F.J. Huyse, J.P.J. Slaets, T. Herzog, A. Lobo, J.S. Lyons, B.C. Opmeer, B. Stein, V. Arolt, N. Balogh, G. Cardoso, P. Fink, M. Rigatelli, R. Van Dijck, G.J. Mellenbergh, Care complexity in the general hospital: Results from a European study, Psychosomatics. (2001). doi:https://doi.org/10.1176/appi.psy.42.3.204.

[60] K. Rockwood, R. Hubbard, Frailty and the geriatrician, Age and Ageing. (2004). doi:https://doi.org/10.1093/ageing/afh153.

[61] J. De Lepeleire, S. Iliffe, E. Mann, J.M. Degryse, Frailty: An emerging concept for general practice, British Journal of General Practice. (2009). doi:https://doi.org/10.3399/bjgp09X420653.

[62] M. Sinha, Spotlight on Canadians: Results from the General Social Survey Portrait of caregivers, 2012, Statistics Canada. (2012).doi

https://doi.org/10.1162/ISEC.

[63] M.J. Faber, R.J. Bosscher, M.J. Chin A Paw, P.C. van Wieringen, Effects of Exercise Programs on Falls and Mobility in Frail and Pre-Frail Older Adults: A Multicenter Randomized Controlled Trial, Archives of Physical Medicine and Rehabilitation. (2006).doi:https://doi.org/10.1016/j.apmr.2006.04.005.

[64] N.K. Latham, C.S. Anderson, A. Lee, D.A. Bennett, A. Moseley, I.D. Cameron, A randomized, controlled trial of quadriceps resistance exercise and vitamin D in frail older people: The frailty interventions trial in elderly subjects (FITNESS), Journal of the American Geriatrics Society. (2003). doi:https://doi.org/10.1046/j.1532-5415.2003.51101.x.

[65] I.Y. Jang, H.W. Jung, H. Park, C.K. Lee, S.S. Yu, Y.S. Lee, E. Lee, R.J. Glynn, D.H. Kim, A multicomponent frailty intervention for socioeconomically vulnerable older adults: A designed-delay study, Clinical Interventions in Aging. (2018).doi:https://doi.org/10.2147/CIA.S177018.

[66] O. Laosa, C. Alonso, M. Castro, L. Rodriguez-Manas, Pharmaceutical Interventions for Frailty and Sarcopenia, Current Pharmaceutical Design. (2014). doi:https://doi.org/10.2174/13816128113196660705.

[67] G. Ellis, P. Langhorne, Comprehensive geriatric assessment for older hospital patients, British Medical Bulletin. (2004). doi:https://doi.org/10.1093/bmb/ldh033.

[68] J. Apóstolo, R. Cooke, E. Bobrowicz-Campos, S. Santana, M. Marcucci, A. Cano, M. Vollenbroek-Hutten, F. Germini, C. Holland, Predicting risk and outcomes for frail older adults: An umbrella review of frailty screening tools, JBI Database of Systematic Reviews and Implementation Reports. (2017).doi:https://doi.org/10.11124/JBISRIR-2016-003018.

[69] R.M. Neal, Pattern Recognition and Machine Learning, Technometrics. (2007). doi:[https://doi.org/10.1198/tech.2007.s518.

[70] A.L. Samuel, Some studies in machine learning using the game of checkers, IBM Journal of Research and Development. (2000). doi:https://doi.org/10.1147/rd.441.0206.

[71] T.M. Mitchell, Machine Learning Machine Learning, Computer. (1997).

[72] Y. Bastanlar, M. ozuysal, Introduction to Machine Learning, in: Methods in Molecular Biology (Clifton, N.J.), 2014: pp. 105–128. doi:https://doi.org/10.1007/978-1-62703-748-8_7.

[73] D. Bzdok, N. Altman, M. Krzywinski, Statistics versus machine learning, Nature Methods. (2018).doi:https://doi.org/10.1038/nmeth.4642.

[74] A.K. Waljee, P.D.R. Higgins, Machine learning in medicine: A primer for physicians, American Journal of Gastroenterology. (2010).doi:https://doi.org/10.1038/ajg.2010.173.

[75] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statiscal Learning with Applications in R, 2013. doi:https://doi.org/10.1016/j.peva.2007.06.006.

[76] M. Kuhn, K. Johnson, Applied predictive modeling, 2013. doi:https://doi.org/10.1007/978-1-4614-6849-3.

[77] D.S. Wilks, Cluster Analysis, 2011. doi:https://doi.org/10.1016/B978-0-12-385022-5.00015-4.

[78] I. De Feis, Dimensionality reduction, in: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 2018. doi:https://doi.org/10.1016/B978-0-12-809633-8.20336-1.

[79] D. Miljkovic, Brief review of self-organizing maps, in: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings, 2017.doi:https://doi.org/10.23919/MIPRO.2017.7973581.

[80] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (2013) 1798–1828.doi:https://doi.org/10.1109/TPAMI.2013.50.

[81] B.W. Silverman, Density estimation: For statistics and data analysis, 2018.doi:https://doi.org/10.1201/9781315140919.

[82] P. Wittek, Quantum Machine Learning article, Quantum Machine Learning. (2014). doi:https://doi.org/10.1016/B978-0-12-800953-6.00015-3.

[83] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, in: International Journal of Data Warehousing and Mining, 2007. doi:https://doi.org/10.4018/jdwm.2007070101.

[84] M.J. Er, R. Venkatesan, W. Ning, An online universal classifier for binary, multi-class and multi-label classification, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings, 2017. doi:https://doi.org/10.1109/SMC.2016.7844809.

[85] M. Naeem Ayyaz, I. Javed, W. Mahmood, Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction, J. Engg. and Appl. Sci. (2012).

[86] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: Closing the gap

to human-level performance in face verification, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014. https://doi.org/10.1109/CVPR.2014.220.

[87] C.W. Hsu, C.J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks. (2002).doi:https://doi.org/10.1109/72.991427.

[88] S.D. Bay, Combining Nearest Neighbor Classifiers Through Multiple Feature Subsets, in: Proceedings of the 17th International Conference on Machine Learning, 1998. doi:https://doi.org/10.1.1.52.455.

[89] B.D. Ripley, Pattern Recognition via Neural Networks, Cambridge University Press, Cambridge. (1996).

[90] S.L. Salzberg, C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993, Machine Learning. 16 (1994) 235–240.doi:https://doi.org/10.1007/BF00993309.

[91] M. Aly, Survey on multiclass classification methods, Neural Netw. (2005).

[92] A.C. Lorena, A.C.P.L.F. De Carvalho, J.M.P. Gama, A review on the combination of binary classifiers in multiclass problems, Artificial Intelligence Review. (2008). doi:https://doi.org/10.1007/s10462-009-9114-9.

[93] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters. (2006).doi:https://doi.org/10.1016/j.patrec.2005.10.010.

[94] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining Multi-label Data, in: Data Mining and Knowledge Discovery Handbook, Springer US, Boston, MA, 2009: pp. 667–685. doi:https://doi.org/10.1007/978-0-387-09823-4_34.

[95] F. Herrera, F. Charte, A.J. Rivera, M.J. Del Jesus, Multilabel classification: Problem analysis, metrics and techniques, 2016. doi:https://doi.org/10.1007/978-3-319-41111-8.

[96] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Dzeroski, An extensive experimental comparison of methods for multi-label learning, in: Pattern Recognition, 2012.doi:https://doi.org/10.1016/j.patcog.2012.03.004.

[97] E. Loza , k Brinker ,Multilabel classification via calibrated label ranking, Machine Learning. (2008).

[98] S. Vembu, T. Gartner , Label ranking algorithm: A survey , in : preference learning , 2011. doi:https://doi.org/10.1007/978-3-642-14125-6_3

[99] J. Read, Scalable Multi-label Classification, University of Waikato, 2010.

[100] E. Gibaja, S. Ventura, A Tutorial on Multilabel Learning, ACM Computing Surveys. 47 (2015) 1–38. doi:https://doi.org/10.1145/2716262.

[101] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine Learning. 85 (2011) 333–359. doi:https://doi.org/10.1007/s10994-011-5256-5.

[102] J. Read, A pruned problem transformation method for multi-label classification, in: New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings, 2008.

[103] G. Tsoumakas, I. Katakis, I. Vlahavas, Effective and efficient multilabel classification in domains with large number of labels, in: Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08), 2008.

[104] A. Clare, R.D. King, Knowledge Discovery in Multi-label Phenotype Data, in: Lecture Notes in Computer Science, 2001: pp. 42–53. doi:https://doi.org/10.1007/3-540-44794-6_4.

[105] R.E. Schapire, Y. Singer, BoosTexter: a boosting-based system for text categorization, Machine Learning. (2000).doi:https://doi.org/10.1023/A:1007649029923.

[106] F. De Comite, R. Gilleron, M. Tommasi, Learning multi-label alternating decision trees from texts and data, in: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), 2003. doi:https://doi.org/10.1007/3-540-45065-3_4.

[107] M.L. Zhang, Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition. (2007). doi:https://doi.org/10.1016/j.patcog.2006.12.019.

[108] M.L. Zhang, Z.H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization, IEEE Transactions on Knowledge and Data Engineering. (2006).doi:https://doi.org/10.1109/TKDE.2006.162.

[109] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, in: Advances in Neural Information Processing Systems 14, The MIT Press, 2002. doi:https://doi.org/10.7551/mitpress/1120.003.0092.

[110] N. Li, Z.H. Zhou, Selective ensemble of classifier chains, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013.doi:https://doi.org/10.1007/978-3-642-38067-9_13.

[111] G. Tsoumakas, I. Katakis, I. Vlahavas, Random k-labelsets for multi-label classification, IEEE Transactions on Knowledge and Data Engineering. (2011).doi:https://doi.org/10.1109/TKDE.2010.164.

[112] J. Read, B. Pfahringer, G. Holmes, Multi-label classification using ensembles of pruned sets, in: Proceedings - IEEE International Conference on Data Mining, ICDM, 2008.doi:https://doi.org/10.1109/ICDM.2008.74.

[113] L. Rokach, A. Schclar, E. Itach, Ensemble methods for multi-label classification, Expert Systems with Applications. (2014). doi:https://doi.org/10.1016/j.eswa.2014.06.015.

[114] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Transactions on Neural Networks. (2005).doi:https://doi.org/10.1109/TNN.2005.845141.

[115] D. Mladenic, Feature selection for dimensionality reduction, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006.doi:https://doi.org/10.1007/11752790_5.

[116] J. Kleinberg, An impossibility theorem for clustering, in: Advances in Neural Information Processing Systems, 2003.

[117] W.S. Sarle, A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Technometrics. 32 (1990) 227. doi:https://doi.org/10.2307/1268876.

[118] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems. (2001).doi:https://doi.org/10.1023/A:1012801612483.

[119] J. Abonyi, B. Feil, Cluster analysis for data mining and system identification, 2007.doi:https://doi.org/10.1007/978-3-7643-7988-9.

[120] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A.Y. Zomaya, S. Foufou, A. Bouras, A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, IEEE Transactions on Emerging Topics in Computing. 2 (2014) 267–279. doi:https://doi.org/10.1109/TETC.2014.2330519.

[121] A. Dharmarajan, T. Velmurugan, Applications of partition based clustering algorithms: A survey, in: 2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013, 2013.doi:https://doi.org/10.1109/ICCIC.2013.6724235.

[122] M. Kuchaki Rafsanjani, Z. Asghari Varzaneh, N. Emami Chukanlo, A Survey Of Hierarchical Clustering Algorithms, Journal of Mathematics and Computer Science. 05 (2012) 229–240.

doi:https://doi.org/10.22436/jmcs.05.03.11.

[123] D.J. Strauss, J.A. Hartigan, Clustering Algorithms, Biometrics. (1975). doi:https://doi.org/10.2307/2529577.

[124] W.-K. Loh, Y.H. Park, A Survey on Density-Based Clustering Algorithms, in: 2014: pp. 775–780. doi:https://doi.org/10.1007/978-3-642-41671-2_98.

[125] R.J.G.B. Campello, D. Moulavi, J. Sander, Density-Based Clustering Based on Hierarchical Density Estimates, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013: pp. 160–172. doi:https://doi.org/10.1007/978-3-642-37456-2_14.

[126] M. Daszykowski, B. Walczak, Density-Based Clustering Methods, in: Comprehensive Chemometrics, 2009. doi:https://doi.org/10.1016/B978-044452701-1.00067-3.

[127] C. Hennig, M. Meila, F. Murtagh, R. Rocci, Handbook of cluster analysis, 2015.doi:https://doi.org/10.1201/b19706.

[128] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics. (2010).doi:https://doi.org/10.1002/wics.101.

[129] H. Johs, H. Johs, Multiple correspondence analysis, in: Multiple Correspondence Analysis For The Social Sciences, 2018. doi:https://doi.org/10.4324/9781315516257-3.

[130] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research. (2008).

[131] A. Lehman, N. O'Rourke, L. Hatcher, E.J. Stepansk, JMP for basic univariate and multivariate statistics: Methods for researchers and social scientists, 2013.

[132] R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules, in: Proc. of 20th International Conference on Very Large Data Bases, VLDB'94, 1994.

[133] T.A. Kumbhare, S. V. Chobe, An Overview of Association Rule Mining Algorithms, International Journal of Computer Science and Information Technologies. (2014).

[134] N. Peek, C. Combi, R. Marin, R. Bellazzi, Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes, Artificial Intelligence in Medicine. (2015).doi:https://doi.org/10.1016/j.artmed.2015.07.003.

[135] C.K. Reddy, C.C. Aggarwal, Healthcare data analytics,

2015.doi:https://doi.org/10.1201/b18588.

[136] K.W. Johnson, J. Torres Soto, B.S. Glicksberg, K. Shameer, R. Miotto, M. Ali, E. Ashley, J.T. Dudley, Artificial Intelligence in Cardiology, Journal of the American College of Cardiology. (2018). doi:https://doi.org/10.1016/j.jacc.2018.03.521.

[137] J. Wiens, E.S. Shenoy, Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology, Clinical Infectious Diseases. (2018). doi:https://doi.org/10.1093/cid/cix731.

[138] E. Commission, White Paper on Artificial Intelligence - A European approach to excellence and trust, 2020.

[139] R. Snyderman, J. Langheier, Prospective health care: The second transformation of medicine, Genome Biology. (2006). doi:https://doi.org/10.1186/gb-2006-7-2-104.

[140] R. Pillai, P. Oza, P. Sharma, Review of Machine Learning Techniques in Health Care, in: 2020: pp. 103–111. doi:https://doi.org/10.1007/978-3-030-29407-6_9.

[141] M. Zhang, P. Wong, Genetic programming for medical classification: A program simplification approach, Genetic Programming and Evolvable Machines. (2008). doi:https://doi.org/10.1007/s10710-008-9059-9.

[142] N. Sut, M. Şenocak, Assessment of the performances of multilayer perceptron neural networks in comparison with recurrent neural networks and two statistical methods for diagnosing coronary artery disease, Expert Systems. (2007).doi:https://doi.org/10.1111/j.1468-0394.2007.00425.x.

[143] F.S. Aguiar, R.C. Torres, J.V.F. Pinto, A.L. Kritski, J.M. Seixas, F.C.Q. Mello, Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil, Medical and Biological Engineering and Computing. (2016).doi:https://doi.org/10.1007/s11517-016-1465-1.

[144] F. Ibrahim, T. Faisal, M.I. Mohamad Salim, M.N. Taib, Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network, Medical and Biological Engineering and Computing. (2010). doi:https://doi.org/10.1007/s11517-010-0669-z.

[145] M. Sangi, K.T. Win, F. Shirvani, M.R. Namazi-Rad, N. Shukla, Applying a novel combination of techniques to develop a predictive model for diabetes complications, PLoS ONE. (2015). doi:https://doi.org/10.1371/journal.pone.0121569.

[146] M. Xu, T.C. Wong, K.S. Chin, Modeling daily patient arrivals at Emergency Department and quantifying the relative importance of contributing variables using artificial neural network, Decision Support Systems. (2013).doi:https://doi.org/10.1016/j.dss.2012.12.019.

[147] M. Suka, S. Oeda, T. Ichimura, K. Yoshida, J. Takezawa, Neural Networks Applied to Medical Data for Prediction of Patient Outcome, in: 2008: pp. 309–325. doi:https://doi.org/10.1007/978-0-387-74935-8_23.

[148] A. Kampouraki, D. Vassis, P. Belsis, C. Skourlas, e-Doctor: A Web based Support Vector Machine for Automatic Medical Diagnosis, Procedia - Social and Behavioral Sciences. (2013). doi:https://doi.org/10.1016/j.sbspro.2013.02.078.

[149] Q. Yan, H. Yan, F. Han, X. Wei, T. Zhu, SVM-based decision support system for clinic aided tracheal intubation predication with multiple features, Expert Systems with Applications. (2009). doi:https://doi.org/10.1016/j.eswa.2008.07.076.

[150] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Medical Informatics and Decision Making. (2011). doi:https://doi.org/10.1186/1472-6947-11-51.

[151] R. Casanova, S. Saldana, E.Y. Chew, R.P. Danis, C.M. Greven, W.T. Ambrosius, Application of random forests methods to diabetic retinopathy classification analyses, PLoS ONE. (2014).doi:https://doi.org/10.1371/journal.pone.0098587.

[152] B. Dai, R.C. Chen, S.-Z. Zhu, W.-W. Zhang, Using Random Forest Algorithm for Breast Cancer Diagnosis, in: 2018 International Symposium on Computer, Consumer and Control (IS3C), IEEE, 2018: pp. 449–452.doi:https://doi.org/10.1109/IS3C.2018.00119.

[153] M.N.M. Garcia, J.C.B. Herraez, M.S. Barba, F.S. Hernandez, Random forest based ensemble classifiers for predicting healthcare-associated infections in intensive care units, in: Advances in Intelligent Systems and Computing, 2016. doi:https://doi.org/10.1007/978-3-319-40162-1_33.

[154] P. Kaur, R. Kumar, M. Kumar, A healthcare monitoring system using random forest and internet of things (IoT), Multimedia Tools and Applications. (2019). doi:https://doi.org/10.1007/s11042-019-7327-8.

[155] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, Decision trees: An overview and their use in medicine, Journal of Medical Systems. (2002). doi:https://doi.org/10.1023/A:1016409317640.

[156] P. Schmitz, LOGISTIC REGRESSION IN MEDICAL DECISION MAKING AND EPIDEMIOLOGY, 1986.

[157] D.G. Kleinbaum, L.L. Kupper, L.E. Chambless, Logistic regression analysis of epidemiologic data: Theory and practice, Communications in Statistics - Theory and Methods. (1982). doi:https://doi.org/10.1080/03610928208828251.

[158] S.C. Bagley, H. White, B.A. Golomb, Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain, Journal of Clinical Epidemiology. (2001). doi:https://doi.org/10.1016/S0895-4356(01)00372-9.

[159] A. Darwish, Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications, Future Computing and Informatics Journal. (2018). doi:https://doi.org/10.1016/j.fcij.2018.06.001.

[160] C. Picardi, J. Cosgrove, S.L. Smith, S. Jamieson, J.E. Alty, Objective Assessment of Cognitive Impairment in Parkinson's Disease Using Evolutionary Algorithm, in: 2017: pp. 109–124.doi:https://doi.org/10.1007/978-3-319-55849-3_8.

[161] C.S. Wang, C.J. Juan, C.C. Yeh, T.Y. Lin, S.Y. Chiang, Prediction model of cervical spine disease established by genetic programming, in: ACM International Conference Proceeding Series, 2017. doi:https://doi.org/10.1145/3092090.3092097.

[162] C. Ryan, K. Krawiec, U.M. O'Reilly, J. Fitzgerald, D. Medernach, Building a stage 1 computer aided detector for breast cancer using genetic programming, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014.

[163] D.B. Neill, Using artificial intelligence to improve hospital inpatient care, IEEE Intelligent Systems. (2013). doi:https://doi.org/10.1109/MIS.2013.51.

[164] C.C. Aggarwal, Data Classification, Chapman and Hall/CRC, 2014.doi:https://doi.org/10.1201/b17320.

[165] M.M.R. Khan, R.B. Arif, A.B. Siddique, M.R. Oishe, Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository, in: 4th International Conference on Electrical Engineering and Information and Communication Technology, ICEEiCT 2018, 2019. doi:https://doi.org/10.1109/CEEICT.2018.8628041.

[166] A.L. Wang, B.X. Chen, C.G. Wang, D.D. Hua, Non-intrusive load monitoring algorithm based on features of V–I trajectory, Electric Power Systems Research. (2018). doi:https://doi.org/10.1016/j.epsr.2017.12.012.

[167] G. Tsoumakas, I. Katakis, Multi-Label Classification, International Journal of Data Warehousing and Mining. 3 (2007) 1–13.doi:https://doi.org/10.4018/jdwm.2007070101.

[168] Z. Ahmadi, S. Kramer, A label compression method for online multi-label classification, Pattern Recognition Letters. 111 (2018) 64–71. doi:https://doi.org/10.1016/j.patrec.2018.04.015.

[169] Y. Zhang, Y. Wang, X.Y. Liu, S. Mi, M.L. Zhang, Large-scale multi-label classification using unknown streaming images: Large-scale multi-label classification using unknown streaming images, Pattern Recognition. (2020).doi:https://doi.org/10.1016/j.patcog.2019.107100.

[170] T.T. Nguyen, M.T. Dang, A.V. Luong, A.W.C. Liew, T. Liang, J. McCall, Multi-label classification via incremental clustering on an evolving data stream, Pattern Recognition. (2019).doi:https://doi.org/10.1016/j.patcog.2019.06.001.

[171] G. Yu, C. Domeniconi, H. Rangwala, G. Zhang, Z. Yu, Transductive multi-label ensemble classification for protein function prediction, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12, ACM Press, New York, New York, USA, 2012: p. 1077.doi:https://doi.org/10.1145/2339530.2339700.

[172] S.C. Dharmadhikari, A Novel Multi label Text Classification Model using Semi supervised learning, International Journal of Data Mining and Knowledge Management Process. 2 (2012) 11–20. doi:https://doi.org/10.5121/ijdkp.2012.2402.

[173] G. Tsoumakas, I. Vlahavas, Random k-Labelsets: An Ensemble Method for Multilabel Classification, in: Machine Learning: ECML 2007, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007: pp. 406–417. doi:https://doi.org/10.1007/978-3-540-74958-5_38.

[174] J. Furnkranz, E. Hullermeier, E. Loza Mencia, K. Brinker, Multilabel classification via calibrated label ranking, Machine Learning. 73 (2008) 133–153. doi:https://doi.org/10.1007/s10994-008-5064-8.

[175] M.-L. Zhang, Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition. 40 (2007) 2038–2048.doi:https://doi.org/10.1016/j.patcog.2006.12.019.

[176] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognition. 37 (2004) 1757–1771. doi:https://doi.org/10.1016/j.patcog.2004.03.009.

[177] Min-Ling Zhang, Zhi-Hua Zhou, A k-nearest neighbor based algorithm for multi-label classification, in: 2005. doi:https://doi.org/10.1109/grc.2005.1547385.

[178] E.L. Mencia, J. Furnkranz, Pairwise learning of multilabel classifications with perceptrons, in: Proceedings of the International Joint Conference on Neural Networks, 2008. doi:https://doi.org/10.1109/IJCNN.2008.4634206.

[179] G. Tsoumakas, I. Katakis, I. Vlahavas, A Review of Multi-Label Classification Methods, Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD 2006). (2006).

[180] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Systems with Applications. 73 (2017) 220–239.doi:https://doi.org/10.1016/j.eswa.2016.12.035.

[181] C.A. Catania, F. Bromberg, C.G. Garino, An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection, in: Expert Systems with Applications, 2012. doi:https://doi.org/10.1016/j.eswa.2011.08.068.

[182] Y.M. Huang, C.M. Hung, H.C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, Nonlinear Analysis: Real World Applications. (2006).doi:https://doi.org/10.1016/j.nonrwa.2005.04.006.

[183] A. Jain, S. Ratnoo, D. Kumar, Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach, in: 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), IEEE, 2017: pp. 1–8.doi:https://doi.org/10.1109/ICOMICON.2017.8279150.

[184] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis. (2002).doi:https://doi.org/10.3233/ida-2002-6504.

[185] N. V. Chawla, N. Japkowicz, A. Kotcz, Special Issue on Learning from Imbalanced Data Sets, ACM SIGKDD Explorations Newsletter. (2004).doi:https://doi.org/10.1145/1007730.1007733.

[186] Haibo He, E.A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering. 21 (2009) 1263–1284.

doi:https://doi.org/10.1109/TKDE.2008.239.

[187] Y. Sun, A.K.C. Wong, M.S. Kamel, Classification of imbalanced data: A review, International Journal of Pattern Recognition and Artificial Intelligence. (2009).doi:https://doi.org/10.1142/S0218001409007326.

[188] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, Y. Zhou, A novel ensemble method for classifying imbalanced data, Pattern Recognition. (2015).doi:https://doi.org/10.1016/j.patcog.2014.11.014.

[189] W.W.Y. Ng, G. Zeng, J. Zhang, D.S. Yeung, W. Pedrycz, Dual autoencoders features for imbalance classification problem, Pattern Recognition. (2016). doi:https://doi.org/10.1016/j.patcog.2016.06.013.

[190] M. Fang, Y. Xiao, C. Wang, J. Xie, Multi-label Classification: Dealing with Imbalance by Combining Labels, in: Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2014. doi:https://doi.org/10.1109/ICTAI.2014.42.

[191] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, Dealing with difficult minority labels in imbalanced mutilabel data sets, Neurocomputing. (2019).doi:https://doi.org/10.1016/j.neucom.2016.08.158.

[192] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, Addressing imbalance in multilabel classification: Measures and random resampling algorithms, Neurocomputing. (2015). doi:https://doi.org/10.1016/j.neucom.2014.08.091.

[193] F. Charte, D. Charte, Working with multilabel datasets in R: The mldr package, R Journal. (2015). doi:https://doi.org/10.32614/rj-2015-027.

[194] F.C. Bernardini, R.B. da Silva, R.M. Rodovalho, E.B.M. Meza, Cardinality and Density Measures and Their Influence to Multi-Label Learning Methods, Learning and Nonlinear Models. 12 (2014) 53–71. doi:https://doi.org/10.21528/LNLM-vol12-no1-art4.

[195] F. Charte, A. Rivera, M.J. del Jesus, F. Herrera, A First Approach to Deal with Imbalance in Multi-label Datasets, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013: pp. 150–160.doi:https://doi.org/10.1007/978-3-642-40846-5_16.

[196] F. Charte, A. Rivera, M.J. del Jesus, F. Herrera, Concurrence among Imbalanced Labels and Its Influence on Multilabel Resampling Algorithms, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014: pp. 110–121.doi:https://doi.org/10.1007/978-3-319-07617-1_10.

[197] X.Y. Liu, J. Wu, Z.H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. (2009). doi:https://doi.org/10.1109/TSMCB.2008.2007853.

[198] F.J. Castellanos, J.J. Valero-Mas, J. Calvo-Zaragoza, J.R. Rico-Juan, Oversampling imbalanced data in the string space, Pattern Recognition Letters. (2018). doi:https://doi.org/10.1016/j.patrec.2018.01.003.

[199] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization, Neurocomputing. 326–327 (2019) 110–122.doi:https://doi.org/10.1016/j.neucom.2017.01.118.

[200] X. Ai, J. Wu, V.S. Sheng, Y. Yao, P. Zhao, Z. Cui, Best First Over-Sampling for Multilabel Classification, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15, ACM Press, New York, New York, USA, 2015: pp. 1803–1806.doi:https://doi.org/10.1145/2806416.2806634.

[201] R. Miranda Pereira, Y. Maldonado E Gomes Da Costa, C.N. Silla, Dealing with imbalanceness in hierarchical multi-label datasets using multi-label resampling techniques, in: Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI, 2018.doi:https://doi.org/10.1109/ICTAI.2018.00128.

[202] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, MLeNN: A First Approach to Heuristic Multilabel Undersampling, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014: pp. 1–9. doi:https://doi.org/10.1007/978-3-319-10840-7_1.

[203] D.L. Wilson, Asymptotic Properties of Nearest Neighbor Rules Using Edited Data, IEEE Transactions on Systems, Man, and Cybernetics. SMC-2 (1972) 408–421. doi:https://doi.org/10.1109/TSMC.1972.4309137.

[204] R.M. Pereira, Y.M.G. Costa, C.N. Silla Jr., MLTL: A multi-label approach for the Tomek Link undersampling algorithm, Neurocomputing. 383 (2020) 95–105. doi:https://doi.org/10.1016/j.neucom.2019.11.076.

[205] I. Tomek, Two Modifications of CNN, IEEE Transactions on Systems, Man, and Cybernetics. SMC-6 (1976) 769–772. doi:https://doi.org/10.1109/TSMC.1976.4309452.

[206] A.F. Giraldo-Forero, J.A. Jaramillo-Garzón, J.F. Ruiz-Muñoz, C.G. Castellanos-Domínguez, Managing Imbalanced Data Sets in Multi-label

Problems: A Case Study with the SMOTE Algorithm, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013: pp. 334–342.doi:https://doi.org/10.1007/978-3-642-41822-8_42.

[207] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research. 16 (2002) 321–357.doi:https://doi.org/10.1613/jair.953.

[208] F. Charte, A.J. Rivera, M.J. del Jesus, F. Herrera, MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation, Knowledge-Based Systems. 89 (2015) 385–397. doi:https://doi.org/10.1016/j.knosys.2015.07.019.

[209] B. Liu, G. Tsoumakas, Synthetic Oversampling of Multi-Label Data based on Local Label Distribution, (2019). doi:http://arxiv.org/abs/1905.00609.

[210] S. Dendamrongvit, P. Vateekul, M. Kubat, Irrelevant attributes and imbalanced classes in multi-label text-categorization domains, Intelligent Data Analysis. 15 (2011) 843–859.doi:https://doi.org/10.3233/IDA-2011-0499.

[211] P. Sadhukhan, S. Palit, Reverse-nearest neighborhood based oversampling for imbalanced, multi-label datasets, Pattern Recognition Letters. 125 (2019) 813–820. doi:https://doi.org/10.1016/j.patrec.2019.08.009.

[212] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 42 (2012) 463–484. doi:https://doi.org/10.1109/TSMCC.2011.2161285.

[213] J. Burez, D. Van den Poel, Handling class imbalance in customer churn prediction, Expert Systems with Applications. 36 (2009) 4626–4636.doi:https://doi.org/10.1016/j.eswa.2008.05.027.

[214] C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, Workshop on Learning from Imbalanced Datasets II. (2003). doi:https://doi.org/10.1.1.68.6858.

[215] K. Chen, B.L. Lu, J.T. Kwok, Efficient classification of multi-label and imbalanced data using min-max modular classifiers, in: IEEE International Conference on Neural Networks - Conference Proceedings, 2006. doi:https://doi.org/10.1109/ijcnn.2006.246893.

[216] B.L. Lu, M. Ito, Task decomposition and module combination

based on class relations: A modular neural network for pattern classification, IEEE Transactions on Neural Networks. (1999). doi:https://doi.org/10.1109/72.788664.

[217] G. Tepvorachai, C. Papachristou, Multi-label imbalanced data enrichment process in neural net classifier training, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008: pp. 1301–1307. doi:https://doi.org/10.1109/IJCNN.2008.4633966.

[218] F.F. Luo, W.Z. Guo, G.L. Chen, Addressing Imbalance in Weakly Supervised Multi-Label Learning, IEEE Access. (2019). doi:https://doi.org/10.1109/ACCESS.2019.2906409.

[219] J. He, H. Gu, W. Liu, Imbalanced Multi-Modal Multi-Label Learning for Subcellular Localization Prediction of Human Proteins with Both Single and Multiple Sites, PLoS ONE. 7 (2012) e37155.doi:https://doi.org/10.1371/journal.pone.0037155.

[220] M.L. Zhang, Ml-rbf: RBF Neural Networks for Multi-Label Learning, Neural Processing Letters. (2009). doi:https://doi.org/10.1007/s11063-009-9095-3.

[221] M.L. Zhang, Z.J. Wang, MIMLRBF: RBF neural networks for multi-instance multi-label learning, Neurocomputing. (2009). doi:https://doi.org/10.1016/j.neucom.2009.07.008.

[222] K.W. Sun, C.H. Lee, Addressing class-imbalance in multi-label learning via two-stage multi-label hypernetwork, Neurocomputing. (2017).doi:https://doi.org/10.1016/j.neucom.2017.05.049.

[223] M.L. Zhang, Y.K. Li, X.Y. Liu, Towards class-imbalance aware multi-label learning, in: IJCAI International Joint Conference on Artificial Intelligence, 2015.

[224] H. Han, M. Huang, Y. Zhang, J. Liu, Decision Support System for Medical Diagnosis Utilizing Imbalanced Clinical Data, Applied Sciences. 8 (2018) 1597. doi:https://doi.org/10.3390/app8091597.

[225] S. Pouyanfar, T. Wang, S.-C. Chen, A Multi-label Multimodal Deep Learning Framework for Imbalanced Data Classification, in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2019: pp. 199–204. doi:https://doi.org/10.1109/MIPR.2019.00043.

[226] K. Sozykin, S. Protasov, A. Khan, R. Hussain, J. Lee, Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks, in: Proceedings - 2018 IEEE/ACIS 19th

International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2018, 2018.doi:https://doi.org/10.1109/SNPD.2018.8441034.

[227] C. Li, G. Shi, Improvement of learning algorithm for the multi-instance multi-label RBF neural networks trained with imbalanced samples, Journal of Information Science and Engineering. (2013).

[228] Y. Xie, D. Li, D. Zhang, H. Shuang, An improved multi-label relief feature selection algorithm for unbalanced datasets, in: Advances in Intelligent Systems and Computing, 2018.doi:https://doi.org/10.1007/978-3-319-69096-4_21.

[229] E.S. Xioufis, M. Spiliopoulou, G. Tsoumakas, I. Vlahavas, Dealing with concept drift and class imbalance in multi-label stream classification, in: IJCAI International Joint Conference on Artificial Intelligence, 2011. doi:https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-266.

[230] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence. (1998).doi:https://doi.org/10.1109/34.667881.

[231] M.A. Tahir, J. Kittler, F. Yan, Inverse random under sampling for class imbalance problem and its application to multi-label classification, Pattern Recognition. 45 (2012) 3738–3750.doi:https://doi.org/10.1016/j.patcog.2012.03.014.

[232] M.A. Tahir, J. Kittler, A. Bouridane, Multilabel classification using heterogeneous ensemble of multi-label classifiers, Pattern Recognition Letters. 33 (2012) 513–523. doi:https://doi.org/10.1016/j.patrec.2011.10.019.

[233] S. Wan, Y. Duan, Q. Zou, HPSLPred: An Ensemble Multi-Label Classifier for Human Protein Subcellular Location Prediction with Imbalanced Source, PROTEOMICS. 17 (2017) 1700262.doi:https://doi.org/10.1002/pmic.201700262.

[234] B.M.N. Arjun Pakrashi, Stacked-MLkNN: A stacking based improvement to Multi-Label k-Nearest Neighbours, in: The 1st International Workshop on Learning with Imbalanced Domains: Theory and Applications, 2017.

[235] G.T. Bin Liu, Making Classifier Chains Resilient to Class Imbalance, in: Proceedings of Machine Learning Research, Beijing (2018), 2018.

[236] M.A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, Improving Multilabel Classification Performance by Using Ensemble of Multi-label

Classifiers, in: 2010: pp. 11–21. doi:https://doi.org/10.1007/978-3-642-12127-2_2.

[237] G.I. Winata, M.L. Khodra, Handling imbalanced dataset in multi-label text categorization using Bagging and Adaptive Boosting, in: 2015 International Conference on Electrical Engineering and Informatics (ICEEI), IEEE, 2015: pp. 500–505. doi:https://doi.org/10.1109/ICEEI.2015.7352552.

[238] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition. 40 (2007) 3358–3378. doi:https://doi.org/10.1016/j.patcog.2007.04.009.

[239] Z.A. Daniels, D.N. Metaxas, Addressing imbalance in multi-label classification using structured hellinger forests, in: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2017.

[240] P. Cao, X. Liu, D. Zhao, O. Zaiane, Cost Sensitive Ranking Support Vector Machine for Multi-label Data Learning, in: Advances in Intelligent Systems and Computing, 2017: pp. 244–255. doi:https://doi.org/10.1007/978-3-319-52941-7_25.

[241] G. Wu, Y. Tian, D. Liu, Cost-sensitive multi-label learning with positive and negative label pairwise correlations, Neural Networks. 108 (2018) 411–423. doi:https://doi.org/10.1016/j.neunet.2018.09.003.

[242] J. V. Tsoumakas, G., Xioufis, E.S., Vilcek, MULAN multi-label dataset repository, (n.d.).http://mulan.sourceforge.net/datasets-mlc.html.

[243] J. Read, P. Reutemann, B. Pfahringer, G. Holmes, MEKA: A multi-label/multi-target extension to WEKA, Journal of Machine Learning Research. 17 (2016).

[244] F. Charte, D. Charte, A. Rivera, M.J. del Jesus, F. Herrera, R Ultimate Multilabel Dataset Repository, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016: pp. 487–499.doi:https://doi.org/10.1007/978-3-319-32034-2_41.

[245] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, I. Vlahavas, MULAN: A Java library for multi-label learning, Journal of Machine Learning Research. (2011).

[246] P. Szymanski, T. Kajdanowicz, Scikit-multilearn: A scikit-based Python environment for performing multi-label classification, Journal of Machine Learning Research. (2019).

[247] S. Godbole, S. Sarawagi, Discriminative Methods for Multi-labeled

Classification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2004: pp. 22–30. doi:https://doi.org/10.1007/978-3-540-24775-3_5.

[248] O. Maimon, L. Rokach, eds., Data Mining and Knowledge Discovery Handbook, Springer US, Boston, MA, 2010. doi:https://doi.org/10.1007/978-0-387-09823-4.

[249] E. Gibaja, S. Ventura, A Tutorial on Multilabel Learning, ACM Computing Surveys. 47 (2015) 1–38. doi:https://doi.org/10.1145/2716262.

[250] J. Read, L. Martino, P.M. Olmos, D. Luengo, Scalable multi-output label prediction: From classifier chains to classifier trellises, Pattern Recognition. 48 (2015) 2096–2109. doi:https://doi.org/10.1016/j.patcog.2015.01.004.

[251] R.B. Pereira, A. Plastino, B. Zadrozny, L.H.C. Merschmann, Correlation analysis of performance measures for multi-label classification, Information Processing and Management. (2018). doi:https://doi.org/10.1016/j.ipm.2018.01.002.

[252] V. Kumar, A.K. Pujari, V. Padmanabhan, V.R. Kagita, Group preserving label embedding for multi-label classification, Pattern Recognition. (2019).doi:https://doi.org/10.1016/j.patcog.2019.01.009.

[253] I. Dimou, M. Zervakis, On the Analogy of Classifier Ensembles With Primary Classifiers: Statistical Performance and Optimality, Journal of Pattern Recognition Research. (2013). doi:https://doi.org/10.13176/11.497.

[254] M.R. Ahmadzadeh, M. Petrou, Use of Dempster-Shafer theory to combine classifiers which use different class boundaries, Pattern Analysis and Applications. (2003). doi:https://doi.org/10.1007/s10044-002-0176-4.

[255] M. Ding, Y. Yang, Multi-label imbalanced classification based on assessments of cost and value, Applied Intelligence. (2018). doi:https://doi.org/10.1007/s10489-018-1156-8.

[256] F. Charte, A. Rivera, M.J. Del Jesus, F. Herrera, Resampling multilabel datasets by decoupling highly imbalanced labels, in: Lecture Notes in Artificial Intelligence, 2015. doi:https://doi.org/10.1007/978-3-319-19644-2_41.

[257] W.H. Organisation, Definition of an older or elderly person, (2010). http://www.who.int/healthinfo/survey/ageingdefnolder/en/index.html (accessed December 25, 2019).

[258] T.A. Comans, N.M. Peel, R.E. Hubbard, A.D. Mulligan, L.C. Gray, P.A. Scuffham, The increase in healthcare costs associated with frailty in older people discharged to a post-acute transition care program, Age and Ageing. (2016). https://doi.org/10.1093/ageing/afv196.

[259] B. Vellas, M. Cesari, J. Li, White Book On Frailty, White Book on Frailty. (2016). https://doi.org/10.14283/jfa.2012.3.

[260] K. Rockwood, M. Andrew, A. Mitnitski, A comparison of two approaches to measuring frailty in elderly people, Journals of Gerontology - Series A Biological Sciences and Medical Sciences. (2007).doi:https://doi.org/10.1093/gerona/62.7.738.

[261] C.Y. Chen, S.C. Wu, L.J. Chen, B.H. Lue, The prevalence of subjective frailty and factors associated with frailty in Taiwan., Archives of Gerontology and Geriatrics. (2010). doi:https://doi.org/10.1016/s0167-4943(10)70012-1.

[262] M.L. Homer, N.P. Palmer, K.P. Fox, J. Armstrong, K.D. Mandl, Predicting Falls in People Aged 65 Years and Older from Insurance Claims, American Journal of Medicine. (2017). doi:https://doi.org/10.1016/j.amjmed.2017.01.003.

[263] F. Bertini, G. Bergami, D. Montesi, G. Veronese, G. Marchesini, P. Pandolfi, Predicting Frailty Condition in Elderly Using Multidimensional Socioclinical Databases, Proceedings of the IEEE. 106 (2018) 723–737. doi:https://doi.org/10.1109/JPROC.2018.2791463.

[264] G.A. Aguayo, A.-F. Donneau, M.T. Vaillant, A. Schritz, O.H. Franco, S. Stranges, L. Malisoux, M. Guillaume, D.R. Witte, Agreement Between 35 Published Frailty Scores in the General Population., American Journal of Epidemiology. 186 (2017) 420–434. https://doi.org/10.1093/aje/kwx061.

[265] B. Santos-Eggimann, N. Sirven, Screening for frailty: Older populations and older individuals, Public Health Reviews. (2016). doi:https://doi.org/10.1186/s40985-016-0021-8.

[266] O. Theou, T.D. Brothers, A. Mitnitski, K. Rockwood, Operationalization of frailty using eight commonly used scales and comparison of their ability to predict all-cause mortality, Journal of the American Geriatrics Society. (2013). doi:https://doi.org/10.1111/jgs.12420.

[267] P. Lee, Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets, International Journal of Environmental Research and Public Health. 11 (2014) 9776–9789. doi:https://doi.org/10.3390/ijerph110909776.

[268] A.B. Parsa, H. Taghipour, S. Derrible, A. (Kouros) Mohammadian, Real-time accident detection: Coping with imbalanced data, Accident Analysis and Prevention. (2019). doi:https://doi.org/10.1016/j.aap.2019.05.014.

[269] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, A. Hussain, Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study, IEEE Access. 4 (2016) 7940–7957. doi:https://doi.org/10.1109/ACCESS.2016.2619719.

[270] B.C. Wallace, K. Small, C.E. Brodley, T.A. Trikalinos, Class Imbalance, Redux, in: 2011 IEEE 11th International Conference on Data Mining, IEEE, 2011: pp. 754–763. doi:https://doi.org/10.1109/ICDM.2011.33.

[271] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinformatics. 14 (2013) 106. doi:https://doi.org/10.1186/1471-2105-14-106.

[272] J. Nayak, B. Naik, H.S. Behera, A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications and Challenges, International Journal of Database Theory and Application. (2015).doi:https://doi.org/10.14257/ijdta.2015.8.1.18.

[273] S. Maji, A.C. Berg, J. Malik, Efficient classification for additive kernel SVMs, IEEE Transactions on Pattern Analysis and Machine Intelligence. (2013). doi:https://doi.org/10.1109/TPAMI.2012.62.

[274] N. Shahid, T. Rappon, W. Berta, Applications of artificial neural networks in health care organizational decision-making: A scoping review, PLoS ONE. (2019).doi:https://doi.org/10.1371/journal.pone.0212356.

[275] Park YS, Lek S. Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. Developments in Environmental Modelling, 2016. doi:https://doi.org/10.1016/B978-0-444-63623-2.00007-4.

[276] J. Lee, Patient-Specific Predictive Modeling Using Random Forests: An Observational Study for the Critically Ill, JMIR Medical Informatics. 5 (2017) e3.doi:https://doi.org/10.2196/medinform.6690.

[277] O. Pauly, Random Forests for Medical Applications, Basic and Applied Ecology. (2003).

[278] J.C. Stoltzfus, Logistic regression: A brief primer, Academic Emergency Medicine. (2011). doi:https://doi.org/10.1111/j.1553-2712.2011.01185.x.

[279] R. Poli, W.B. Langdon, N.F. McPhee, A Field Guide to Genetic Programing, 2008.

[280] L. Vanneschi, A. Farinaccio, G. Mauri, M. Antoniotti, P. Provero, M. Giacobini, A comparison of machine learning techniques for survival prediction in breast cancer., BioData Mining. 4 (2011) 12.doi:https://doi.org/10.1186/1756-0381-4-12.

[281] HEAL., HeuristicLab, (2002). https://dev.heuristiclab.com/trac.fcgi/.

[282] A.C. Muller, S. Guido, Introduction to Machine Learning with Python and Scikit-Learn, 2015.

[283] M. Alirezanejad, R. Enayatifar, H. Motameni, H. Nematzadeh, Heuristic filter feature selection methods for medical datasets., Genomics. (2019). doi:https://doi.org/10.1016/j.ygeno.2019.07.002.

[284] D.H. Wolpert, The Lack of a Priori Distinctions between Learning Algorithms, Neural Computation. (1996). doi:https://doi.org/10.1162/neco.1996.8.7.1341.

[285] F. Azimlu, S. Rahnamayan, M. Makrehchi, N. Kalra, Comparing Genetic Programming with Other Data Mining Techniques on Prediction Models, in: 2019 14th International Conference on Computer Science and Education (ICCSE), IEEE, 2019: pp. 785–791. doi:https://doi.org/10.1109/ICCSE.2019.8845381.

[286] C.A. Bannister, J.P. Halcox, C.J. Currie, A. Preece, I. Spasić, A genetic programming approach to development of clinical prediction models: A case study in symptomatic cardiovascular disease, PLoS ONE. (2018). doi:https://doi.org/10.1371/journal.pone.0202685.

[287] B. Can, C. Heavey, A comparison of genetic programming and artificial neural networks in metamodeling of discrete-event simulation models, Computers and Operations Research. (2012). doi:https://doi.org/10.1016/j.cor.2011.05.004.

[288] R. Poli, J. Koza, Genetic programming, in: Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques, Second Edition, 2014. doi:https://doi.org/10.1007/978-1-4614-6940-7_6.

[289] P. Hanlon, B.I. Nicholl, B.D. Jani, D. Lee, R. McQueenie, F.S. Mair, Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality, The Lancet Public Health. (2018). doi:https://doi.org/10.1016/S2468-2667(18)30091-4.

[290] Y.Y. Ding, J. Kuha, M. Murphy, Multidimensional predictors of physical frailty in older people: identifying how and for whom they exert their effects, Biogerontology. (2017). doi:https://doi.org/10.1007/s10522-017-9677-9.

[291] E. Dent, P. Kowal, E.O. Hoogendijk, Frailty measurement in research and clinical practice: A review, European Journal of Internal Medicine. 31 (2016) 3–10. doi:https://doi.org/10.1016/j.ejim.2016.03.007.

[292] R.A. Heuberger, The Frailty Syndrome: A Comprehensive Review, Journal of Nutrition in Gerontology and Geriatrics. 30 (2011) 315–368. doi:https://doi.org/10.1080/21551197.2011.623931.

[293] D.B. Hogan, C. MacKnight, H. Bergman, C.I. on F. and A. Steering Committee, Models, definitions, and criteria of frailty., Aging Clinical and Experimental Research. 15 (2003) 1–29. doi:https://doi.org/10.1016/B978-012369391-4/50051-5.

[294] S. Karunananthan, C. Wolfson, H. Bergman, F. Béland, D.B. Hogan, A multidisciplinary systematic literature review on frailty: Overview of the methodology used by the Canadian Initiative on Frailty and Aging, BMC Medical Research Methodology. (2009). doi:https://doi.org/10.1186/1471-2288-9-68.

[295] M. Ruiz, C. Cefalu, T. Reske, Frailty syndrome in geriatric medicine, American Journal of the Medical Sciences. (2012). doi:https://doi.org/10.1097/MAJ.0b013e318256c6aa.

[296] J.E. Morley, B. Vellas, G. Abellan van Kan, S.D. Anker, J.M. Bauer, R. Bernabei, M. Cesari, W.C. Chumlea, W. Doehner, J. Evans, L.P. Fried, J.M. Guralnik, P.R. Katz, T.K. Malmstrom, R.J. McCarter, L.M. Gutierrez Robledo, K. Rockwood, S. von Haehling, M.F. Vandewoude, J. Walston, Frailty Consensus: A Call to Action, Journal of the American Medical Directors Association. 14 (2013) 392–397.doi:https://doi.org/10.1016/j.jamda.2013.03.022.

[297] P.D. Sloane, M. Cesari, Research on Frailty: Continued Progress, Continued Challenges, Journal of the American Medical Directors Association. 19 (2018) 279–281. doi:https://doi.org/10.1016/j.jamda.2018.01.003.

[298] J. Marcusson, M. Nord, H.J. Dong, J. Lyth, Clinically useful prediction of hospital admissions in an older population, BMC Geriatrics. (2020). doi:https://doi.org/10.1186/s12877-020-1475-6.

[299] M. Khezrian, P.K. Myint, C. McNeil, A.D. Murray, A Review of Frailty Syndrome and Its Physical, Cognitive and

Emotional Domains in the Elderly, Geriatrics. 2 (2017) 36. doi:https://doi.org/10.3390/geriatrics2040036.

[300] H. Bergman, L. Ferrucci, J. Guralnik, D.B. Hogan, S. Hummel, S. Karunananthan, C. Wolfson, Frailty: An emerging research and clinical paradigm - Issues and controversies, Journals of Gerontology - Series A Biological Sciences and Medical Sciences. (2007). doi:https://doi.org/10.1093/gerona/62.7.731.

[301] E. Gibaja, S. Ventura, Multi-label learning: a review of the state of the art and ongoing research, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 4 (2014) 411–444. doi:https://doi.org/10.1002/widm.1139.

[302] A. Tarekegn, F. Ricceri, G. Costa, E. Ferracin, M. Giacobini, Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches., JMIR Medical Informatics. 8 (2020) e16678. doi:https://doi.org/10.2196/16678.

[303] G. Tsoumakas, I. Vlahavas, Random k-labelsets: An ensemble method for multilabel classification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2007.doi:https://doi.org/10.1007/978-3-540-74958-5_38.

[304] E. Hullermeier, J. Fürnkranz, W. Cheng, K. Brinker, Label ranking by learning pairwise preferences, Artificial Intelligence. (2008). doi:https://doi.org/10.1016/j.artint.2008.08.002.

[305] N. Thai-Nghe, Learning optimal threshold on resampling data to deal with class imbalance, in: Proc. IEEE RIVF . . . , 2010.

[306] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explorations Newsletter. (2004). doi:https://doi.org/10.1145/1007730.1007735.

[307] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, X. Geng, Binary relevance for multi-label learning: an overview, Frontiers of Computer Science. 12 (2018) 191–202. doi:https://doi.org/10.1007/s11704-017-7031-7.

[308] F. Markatopoulou, V. Mezaris, I. Kompatsiaris, A comparative study on the use of multi-label classification techniques for concept-based video indexing and annotation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014.doi:https://doi.org/10.1007/978-3-319-04114-8_1.

[309] R. Xu, D. WunschII, Survey of Clustering Algorithms, IEEE Transactions on Neural Networks. 16 (2005) 645–678.doi:https://doi.org/10.1109/TNN.2005.845141.

[310] I. Dokmanic, R. Parhizkar, J. Ranieri, M. Vetterli, Euclidean Distance Matrices: Essential theory, algorithms, and applications, IEEE Signal Processing Magazine. (2015).doi:https://doi.org/10.1109/MSP.2015.2398954.

[311] R. Cordeiro De Amorim, B. Mirkin, Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering, Pattern Recognition. (2012). doi:https://doi.org/10.1016/j.patcog.2011.08.012.

[312] L. Sahu, B.R. Mohan, An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop, in: 9th International Conference on Industrial and Information Systems, ICIIS 2014, 2015.doi:https://doi.org/10.1109/ICIINFS.2014.7036661.

[313] S. Chakraborty, S. Das, K-Means clustering with a new divergence-based distance metric: Convergence and performance analysis, Pattern Recognition Letters. (2017). doi:https://doi.org/10.1016/j.patrec.2017.09.025.

[314] V. Estivill-Castro, Why so many clustering algorithms, ACM SIGKDD Explorations Newsletter. (2002). doi:https://doi.org/10.1145/568574.568575.

[315] E.A. Tanaka, S.R. Nozawa, A.A. Macedo, J.A. Baranauskas, A multi-label approach using binary relevance and decision trees applied to functional genomics, Journal of Biomedical Informatics. (2015). doi:https://doi.org/10.1016/j.jbi.2014.12.011.

[316] P.-N. Tan, M. Steinbach, V. Kumar : Cluster Analysis : Basic Concepts and Algorithms, Introduction to Data Mining. (2005).doi:https://doi.org/10.1016/0022-4405(81)90007-8.

[317] W.M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association. (1971). doi:https://doi.org/10.1080/01621459.1971.10482356.

[318] N.X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, Journal of Machine Learning Research. (2010).

[319] E. Rendon, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus External cluster validation indexes, International Journal of Computers and Communications. (2011).

[320] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics. 20 (1987) 53–65.doi:https://doi.org/10.1016/0377-0427(87)90125-7.

[321] T. Calinski, J. Harabasz, A Dendrite Method Foe Cluster Analysis, Communications in Statistics. (1974). doi:https://doi.org/10.1080/03610927408827101.

[322] D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence. (1979).doi:https://doi.org/10.1109/TPAMI.1979.4766909.

[323] D. Moulavi, P.A. Jaskowiak, R.J.G.B. Campello, A. Zimek, J. Sander, Density-Based Clustering Validation, in: Proceedings of the 2014 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014: pp. 839–847. https://doi.org/10.1137/1.9781611973440.96.

[324] A. Rakhlin, A. Caponnetto, Stability of K-means clustering, in: Advances in Neural Information Processing Systems, 2007. doi:https://doi.org/10.1007/978-3-540-72927-3_4.

[325] J. Wang, Consistent selection of the number of clusters via crossvalidation, Biometrika. (2010).doi:https://doi.org/10.1093/biomet/asq061.

[326] R. Tibshirani, G. Walther, Cluster validation by prediction strength, Journal of Computational and Graphical Statistics. (2005). doi:https://doi.org/10.1198/106186005X59243.

[327] G.W. Milligan, M.C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika. (1985). doi:https://doi.org/10.1007/BF02294245.

[328] A.K. Jain, R.C. Dubes, Clustering Methods and Algorithms, in: Algorithms for Clustering Data, 1988.

[329] S. Chakraborty, D. Paul, S. Das, J. Xu, Entropy Regularized Power k-Means Clustering, (2020). doi:http://arxiv.org/abs/2001.03452.

[330] D.M. Witten, R. Tibshirani, A framework for feature selection in clustering, Journal of the American Statistical Association. (2010). doi:https://doi.org/10.1198/jasa.2010.tm09415.

[331] P. Olukanmi, F. Nelwamondo, T. Marwala, Rethinking k-means clustering in the age of massive datasets: a constant-time approach, Neural Computing and Applications. (2019).doi:https://doi.org/10.1007/s00521-019-04673-0.

[332] Z.C. Goran Petrovic, Comparison of Clustering Methods for Failure

Data Analysis: A Real Life Application, in: Proceedings of the XV International Scientific Conference on Industrial Systems (IS'11), 2011: pp. 297–300.

[333] M. Hassani, T. Seidl, Using internal evaluation measures to validate the quality of diverse stream clustering algorithms, Vietnam Journal of Computer Science. (2017). doi:https://doi.org/10.1007/s40595-016-0086-9.

[334] X. Zhang, H. Zhao, S. Zhang, R. Li, A novel deep neural network model for multi-label chronic disease prediction, Frontiers in Genetics. (2019). doi:https://doi.org/10.3389/fgene.2019.00351.

[335] K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas, Multi-label classification of music into emotions, in: ISMIR 2008 - 9th International Conference on Music Information Retrieval, 2008.

[336] D. Napoleon, S. Pavalakodi, A New Method for Dimensionality Reduction Using KMeans Clustering Algorithm for High Dimensional Data Set, International Journal of Computer Applications. 13 (2011) 41–46. doi:https://doi.org/10.5120/1789-2471.

[337] W. Li, J.E. Cerise, Y. Yang, H. Han, Application of t-SNE to human genetic data, Journal of Bioinformatics and Computational Biology. 15 (2017) 1750017. doi:https://doi.org/10.1142/S0219720017500172.

[338] J. Jin, W. Wang, Influential features PCA for high dimensional clustering, Annals of Statistics. (2016).doi:https://doi.org/10.1214/15-AOS1423.

[339] C. Syms, Principal components analysis, in: Encyclopedia of Ecology, 2018.doi:https://doi.org/10.1016/B978-0-12-409548-9.11152-2.

[340] J.H. Do, D.K. Choi, Normalization of microarray data: Single-labeled and dual-labeled arrays, Molecules and Cells. (2006).

[341] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters. 31 (2010) 651–666. doi:https://doi.org/10.1016/j.patrec.2009.09.011.

[342] Y. Lv, T. Ma, M. Tang, J. Cao, Y. Tian, A. Al-Dhelaan, M. Al-Rodhaan, An efficient and scalable density-based clustering algorithm for datasets with complex structures, Neurocomputing. (2016).doi:https://doi.org/10.1016/j.neucom.2015.05.109.

[343] Lewis, Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting, Butterworth Scientific. (1982). doi:https://doi.org/10.1002/for.3980010202.

[344] S.I. Vuik, E.K. Mayer, A. Darzi, Patient Segmentation Analysis Offers

Significant Benefits For Integrated Care And Support, Health Affairs. 35 (2016) 769–775. doi:https://doi.org/10.1377/hlthaff.2015.1311.

[345] J.L. Chong, D.B. Matchar, Benefits of population segmentation analysis for developing health policy to promote patient-centred care, Annals of the Academy of Medicine Singapore. (2017).

[346] Y. Zhou, Improving Care for Older Adults: A Model to Segment the Senior Population, The Permanente Journal. (2014) 18–21. doi:https://doi.org/10.7812/TPP/14-005.

[347] S. Yan, Y.H. Kwan, C.S. Tan, J. Thumboo, L.L. Low, A systematic review of the clinical application of data-driven population segmentation analysis, BMC Medical Research Methodology. 18 (2018) 121. doi:https://doi.org/10.1186/s12874-018-0584-9.

[348] R. Ruby-Figueroa, Principal Component Analysis (PCA), in: Encyclopedia of Membranes, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015: pp. 1–2.doi:https://doi.org/10.1007/978-3-642-40872-4_1999-1.

[349] R.T. Clarke, M.J. Greenacre, Theory and Applications of Correspondence Analysis, The Journal of Animal Ecology. 54 (1985) 1031. doi:https://doi.org/10.2307/4399.

[350] N. Shi, X. Liu, Y. Guan, Research on k-means clustering algorithm: An improved k-means clustering algorithm, in: 3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI 2010, 2010.doi:https://doi.org/10.1109/IITSI.2010.74.

[351] M. Deliu, M. Sperrin, D. Belgrave, A. Custovic, Identification of Asthma Subtypes Using Clustering Methodologies, Pulmonary Therapy. 2 (2016) 19–41. doi:https://doi.org/10.1007/s41030-016-0017-z.

[352] P.D. McNicholas, Model-Based Clustering, Journal of Classification. 33 (2016) 331–373. doi:https://doi.org/10.1007/s00357-016-9211-9.

[353] R.J.G.B. Campello, D. Moulavi, A. Zimek, J. Sander, Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection, ACM Transactions on Knowledge Discovery from Data. 10 (2015) 1–51. doi:https://doi.org/10.1145/2733381.

[354] M. Pikoula, J.K. Quint, F. Nissen, H. Hemingway, L. Smeeth, S. Denaxas, Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records, BMC Medical Informatics and Decision Making. 19 (2019) 86. doi:https://doi.org/10.1186/s12911-019-0805-0.

[355] O. Chikumbo, V. Granville, Optimal Clustering and Cluster Identity in Understanding High-Dimensional Data Spaces with Tightly Distributed Points, Machine Learning and Knowledge Extraction. 1 (2019) 715–744. doi:https://doi.org/10.3390/make1020042.

[356] M.A. Masud, J.Z. Huang, C. Wei, J. Wang, I. Khan, M. Zhong, I-nice: A new approach for identifying the number of clusters and initial cluster centres, Information Sciences. 466 (2018) 129–151.doi:https://doi.org/10.1016/j.ins.2018.07.034.

[357] Y. Roh, G. Heo, S.E. Whang, A survey on data collection for machine learning: A big data - AI integration perspective, ArXiv. (2018).

[358] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, Npj Computational Materials. (2019). https://doi.org/10.1038/s41524-019-0221-0.

[359] L. Baier, F. Jöhren, S. Seebacher, Challenges in the deployment and operation of machine learning in practice, in: 27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019, 2020.