Performance Comparison of Multi-Label Learning Algorithms on Clinical Data for Chronic Diseases

D. Zufferey[a,b,1,*], T. Hofer[a], J. Hennebert[b], M. Schumacher[a], R. Ingold[b], S. Bromuri[a]

[a] *AiSlab Group, Institute of Information Systems, University of Applied Sciences and Arts Western Switzerland, Techno-Pôle 3, 3960 Sierre, Switzerland*
[b] *DIVA Group, Department of Informatics, University of Fribourg, Bd de Pérolles 90, 1700 Fribourg, Switzerland*

Abstract

We are motivated by the issue of classifying diseases of chronically ill patients to assist physicians in their everyday work. Our goal is to provide a performance comparison of state-of-the-art multi-label learning algorithms for the analysis of multivariate sequential clinical data from medical records of patients affected by chronic diseases. As a matter of fact, the multi-label learning approach appears to be a good candidate for modeling overlapped medical conditions, specific to chronically ill patients. With the availability of such comparison study, the evaluation of new algorithms should be enhanced.

According to the method, we choose a summary statistics approach for the processing of the sequential clinical data, so that the extracted features maintain an interpretable link to their corresponding medical records. The

*Corresponding author

Email addresses: damien.zufferey@hevs.ch (D. Zufferey), thomas.hofer@hevs.ch (T. Hofer), jean.hennebert@unifr.ch (J. Hennebert), michael.schumacher@hevs.ch (M. Schumacher), rolf.ingold@unifr.ch (R. Ingold), stefano.bromuri@hevs.ch (S. Bromuri)

[1] Permanent email address: damien.zufferey@gmail.com

publicly available MIMIC-II dataset, which contains more than 19,000 patients with chronic diseases, is used in this study. For the comparison we selected the following multi-label algorithms: ML-kNN, AdaBoostMH, binary relevance, classifier chains, HOMER and RAkEL.

Regarding the results, binary relevance approaches, despite their elementary design and their independence assumption concerning the chronic illnesses, perform optimally in most scenarios, in particular for the detection of relevant diseases. In addition, binary relevance approaches scale up to large dataset and are easy to learn. However, the RAkEL algorithm, despite its scalability problems when it is confronted to large dataset, performs well in the scenario which consists of the ranking of the labels according to the dominant disease of the patient.
Keywords: multi-label learning, complex patient, chronic disease, clinical data, summary statistics.

---

1. Introduction

Chronic diseases, also called noncommunicable diseases (NCDs) [1], are characterized by a long duration and generally a slow progression. Widespread chronic diseases include: cardiovascular diseases, chronic respiratory diseases and diabetes. Chronic conditions are a major concern for public health programs of governments, particularly due to their negative effect in the continuous growth of medical care costs [2]. Chronic obstructive pulmonary disease (COPD) is an incurable illness, mainly due to tobacco smoking, where the treatment merely slows the progress of the condition. The World Health Organization (WHO) estimates that 64 million people have COPD worldwide

in 2004 [3]. Concerning another major chronic disease, diabetes affects 347 million people worldwide in 2008 [4]. WHO projects that diabetes will be the 7th leading cause of death in 2030 [5]. Type 2 diabetes consists of 90% of people with diabetes, and is mostly the consequence of excess body weight and physical inactivity [6].

Despite the technical progress in the medical area which allows patients to be monitored in a more continuous way [7], the treatment of chronically ill patients, which can develop several comorbidities, remains complex for the physician. The continuous monitoring generates larger quantity of data. Often these measures are heterogeneous, such as: laboratory tests, physiological values or electrocardiograms. On the one side, physicians willing to take optimal decisions will have to aggregate the information contained in these data. On the other side, such aggregation will become (or are already) unmanageable for humans. In addition, physicians are frequently in charge of hundreds of patients, as reported in [8]. Therefore, there is a need for state-of-the-art data-mining and machine learning tools to assist physicians by providing aggregated information about their patients. Indeed, as reported in [9], medical doctors would use tools that improve their understanding of an illness even if these involve more cognitive effort than in the standard practice. Several challenges appear during the design of such tools. Chronically ill patients, such as a diabetic patient, suffer frequently from several comorbidities in relation with the main disease. New approaches in the machine learning field, such as Multi-Label Learning (MLL), which has received, in the last few years, substantial contributions from the machine learning community [10, 11, 12], is then the good candidate for modeling the profile of

a patient affected by several comorbidities. Another challenge concerns the characteristics of medical signals. Clinical data consist of multivariate time series that are often irregular by the fact that a patient may present various number of records with respect to another patient and the values can be nonuniformly sampled. The processing of data with these characteristics is challenging and techniques for the extraction of features are needed. One approach consists on relying on quantization methods, such as k-means clustering and Bag-of-Words (BoW), that have been proven successful in several medical data processing tasks [13]. Another approach would be to extract summary statistics for the different types of sequential clinical data [14].

MLL differs from classical machine learning by tackling the learning problem from a different perspective. In contrast to the classical classification tasks where each observation belongs to only one mutually exclusive class, in MLL decision areas of labels (i.e. classes) overlap. This aspect leads to the annotation (i.e. instead of classification) of observations with zero, one or several labels. In addition, instead of expressing the presence or the absence of a label as a binary variable, it is possible to express the confidence of the presence of a label through a score or a probability. This formulation looks natural for many problems in real life, such as: the detection of emotions in music [15, 16], the semantic scene classification [17] or the classification of text into topics [18].

Regarding the application of such approaches in the medical domains, we can mention several research works. In genomics field, Barutcuoglu et al. proposed a Bayesian framework for the prediction of gene function [19]. Independently for each gene function, a Support Vector Machine (SVM) is

trained, then a Bayesian network is built for combining the multiple classifier results. The graph structure of the network is based on a hierarchical gene taxonomy. The aim of this network is to avoid inconsistent set of predictions, where for a given gene a specific label may be predicted relevant while its inclusive parent label is predicted irrelevant. In the biology field, Xiao et al. developed the iLoc-Virus predictor [20] for predicting the subcellular locations of proteins according to their sequence information. In their work, they focus on viral proteins, those generated by viruses. Being able to predict the locations of viral proteins in a viral infected cell is important for improving antiviral drugs. As a virus protein can have more than one location, MLL methods accommodate well, and thus the ML-kNN [21] algorithm was chosen for their predictor. The following work are focused on chronic diseases, although they are not based on MLL but on related techniques. Huang et al. proposed a system for the prognosis and the diagnosis of chronic diseases which is based on data mining and case-based reasoning [22]. Data mining techniques are used to discover patterns from health examination data. More precisely, a decision tree induction algorithm is applied to find rules which will serve to the chronic diseases classification of new cases. Afterwards, case-based reasoning, which consists on the analysis of old cases to provide solution for a new case, aims to support physicians for the diagnosis and the treatments of chronic diseases. Regarding the evaluation, the experiment data were collected from a professional health examination center, and a feasibility test was performed with 12 discharged real cases. Amaral et al. developed a clinical decision support system to assess patients affected by chronic obstructive pulmonary disease (COPD) based on the forced os-

cillation technique (FOT) [23]. FOT is a noninvasive method to asses the breathing mechanics, using small amplitude pressure oscillations to stimulate the respiratory system in order to evaluate the flow response. Several machine learning classifiers were attempted, such as naive Bayes (NB), k-nearest neighbors (KNN), decision trees (DT), artificial neural networks (ANN), or support vector machines (SVM). Based on a dataset of 50 volunteers (where 25 have COPD), non-linear classifiers such as ANN and SVM and the lazy learning KNN classifier were able to reach a proper accuracy for COPD clinical diagnosis (sensitivity $> 87\%$, specificity $> 94\%$).

We are motivated by the problem of studying multi-label learning techniques for the analysis of clinical data in order to identify patients that may be affected by chronic diseases. We use the MIMIC-II clinical database [24] where 19,773 patients of various intensive care units (ICUs) are diagnosed with one or several chronic diseases according to the coding scheme of the International Classification of Disease revision 9 (ICD-9)[2]. Being able to characterize patients, based on their clinical data, open several applications, such as the identification of patient cohorts in the context of comparative effectiveness studies or in the case of clinical decision support systems [14]. In a previous article [25], Bromuri et al. report on a new classifier which combines BoW and supervised dimensionality reduction algorithms to perform multi-label classification on health records of chronically ill patients. In the framework of this research, we discovered the following new challenges. Although the quantization method (BoW) used is convenient for the feature

[2]http://www.who.int/classifications/icd

extraction when dealing with irregular time series, we think that a finer feature extraction approach based on summary statistics [14] will improve the results while making it easier to identify the influent characteristics.

In addition, the evaluation of a new MLL technique for the classification of chronic diseases based on the analysis of clinical data is made difficult by the fact that there are no studies which provide a large experimental comparison of state-of-the-art MLL algorithms on such data. The main contribution of this work is a large experimental review of multi-label learning approaches for the analysis of clinical data of chronically ill patients. We provide an extended description on properties of the dataset, on the way features are extracted using summary statistics and how the evaluation is conducted.

The rest of this document is organized as follows: Section 2 presents a background on evaluation metrics and methods for multi-label learning; Section 3 describes the MIMIC-II database and its properties; Section 4 defines the methodology for building models; Section 5 presents the results for the multi-label algorithms considered in this study; finally, Section 6 concludes this article and draws the lines for future work.

2. Background

This section begins with the formal definition of a MLL problem and their related evaluation metrics. Then, a state-of-the-art of the existing MLL techniques is described.

With L for the finite set of labels, and with X for the domain of observation, the training set T is defined as $T = \{(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)\}$ ($x_i \in X, Y_i \subseteq L$). Based on these definitions, a multi-label classifier h is defined as

$h : X \to 2^L$. In addition, some evaluation metrics are based on the output of a real-valued scoring function f defined as $f : X \times L \to \mathbb{R}$. For an observation x_i and its attached label set Y_i, the scoring function f will output larger values for labels in Y_i than those not in Y_i, i.e. $f(x_i, y_1) > f(x_i, y_2)$ for any $y_1 \in Y_i$ and $y_2 \notin Y_i$. Finally, some evaluation metrics need also a ranking function $rank_f(\cdot, \cdot)$, which maps the outputs of $f(x_i, y)$ for any $y \in L$ to $\{1, 2, ..., |L|\}$ such that if $f(x_i, y_1) > f(x_i, y_2)$ then $rank_f(x_i, y_1) < rank_f(x_i, y_2)$.

2.1. Evaluation metrics

In classic learning approach of multiclass problems, the evaluation is done through common metrics such as accuracy, precision, and recall. In multi-label problems, the evaluation is more complicated and need extended evaluation metrics. The following five evaluation metrics [21], that are described below, are commonly used with multi-label problems. Note that these evaluation metrics consider all the set of labels, this is not a per label evaluation, thus the results are sensitive to the distribution of labels in the dataset.

Let a testing set $S = \{(x_1, Y_1), (x_2, Y_2), ..., (x_m, Y_m)\}$.

2.1.1. Hamming loss

Hamming loss is defined as the fraction of the proper labels to the total number of labels. The score lies between 0 and 1, where 0 corresponds to the best result:

$$hloss_S(h) = \frac{1}{m} \sum_{i=1}^{m} \frac{|h(x_i) \triangle Y_i|}{|L|}, \tag{1}$$

where $\triangle$ represents the symmetric difference.

2.1.2. One-error

One-error evaluates the fraction of top-ranked labels not part of the relevant label set. The score lies between 0 and 1, where 0 corresponds to the best result:

$$one\text{-}error_S(f) = \frac{1}{m}\sum_{i=1}^{m}\gamma(\arg\max_{y\in L} f(x_i, y)), \tag{2}$$

where

$$\gamma(y) = \begin{cases} 1 & \text{if } y \notin Y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

2.1.3. Coverage

Coverage evaluates how many steps are needed, on the average, to go down the list of labels so as to cover all the relevant labels of the observation. A score as small as possible is better:

$$coverage_S(f) = \frac{1}{m}\sum_{i=1}^{m}\max_{y\in Y_i} rank_f(x_i, y) - 1. \tag{4}$$

2.1.4. Ranking loss

Ranking loss evaluates the average part of reversely ordered label pairs, for the observation. The score lies between 0 and 1, where 0 corresponds to the best result:

$$\begin{aligned} rloss_S(f) = \frac{1}{m}\sum_{i=1}^{m} & \frac{1}{|Y_i||(L \setminus Y_i)|} \\ & \times |\{(y_1, y_2)|f(x_i, y_1) \leqslant f(x_i, y_2), (y_1, y_2) \in Y_i \times (L \setminus Y_i)\}|, \end{aligned} \tag{5}$$

where $\setminus$ is the set-theoretic difference.

2.1.5. Average precision

Average precision evaluates the average fraction of relevant labels ranked above a particular label $y \in Y_i$. The score lies between 0 and 1, where 1 corresponds to the best result:

$$avgprec_S(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|} \times \sum_{y\in Y_i}\frac{|\{y'|rank_f(x_i,y') \leqslant rank_f(x_i,y), y' \in Y_i\}|}{rank_f(x_i,y)}. \quad (6)$$

2.2. Multi-label learning

Madjarov et al. did an extensive comparison of multi-label learning methods on various benchmark datasets [11]. They propose three categories of methods for multi-label learning: algorithm adaptation methods, problem transformation methods and ensemble methods.

2.2.1. Algorithm adaptation methods

Algorithm adaptation methods are modifications of existing machine learning algorithms for solving multi-label learning problems. The modification allows the new algorithm to handle directly multi-label data. Adaptations for multi-label learning have been proposed for: AdaBoost [26], k-nearest neighbors [21], decision trees [27, 28], neural networks [29] and support vector machines [30].

2.2.2. Problem transformation methods

Problem transformation methods transform a multi-label learning problem into several single-label classification problems which allow the use of existing machine learning algorithms. Problem transformation methods can be

divided into three groups: binary relevance methods, label power-set methods and pair-wise methods.

The binary relevance (BR) method [31] splits the multi-label learning problem into several binary classification problems using the one-against-all strategy. Since the BR method has as major drawback to not consider correlation between labels, the classifier chain (CC) approach [32] was introduced as an extension. In a similar way to the BR method, the CC approach involves a binary transformation for each label, and the extension concerns the addition in the feature space of binary variables for label relevances of all previous classifiers, thus resulting to a classifiers chain which considers a form of correlation between labels.

The label power-set (LP) method [31] transforms the multi-label learning problem into a single-class classification problem by taking into account the set of all possible combinations of labels, thus the LP method directly considers the labels correlation. As the space of possible classes can be very large, LP cannot be applied as it is for practical applications. To solve this issue, the Pruned Problem Transformation (PPT) method [33] has been proposed. PPT is an efficient method that keeps, based on the occurrence of observations in the training set, only combined labels that occur more than a predefined threshold, discarding other combinations. Hierarchy Of Multi-label Learners (HOMER) [34] is another approximation of LP which is based on the divide-and-conquer paradigm where the set of labels is organized following a tree structure, build using a clustering algorithm, where each node is a simpler multi-label classification problem dealing only with a subset of labels. This is an efficient algorithm that is adapted to problems containing

large set of labels.

The pair-wise method handles the multi-label learning problem using binary classifiers in a round-robin approach. The idea is to have $L(L-1)/2$ classifiers which consider all pairs of labels. Given an observation, each binary classifier predicts one of the two labels. When all classifiers have been evaluated, a majority voting algorithm is used and the labels are ranked according to their number of votes. Recent works in pair-wise method are [35, 36].

2.2.3. Ensemble methods

Ensemble methods are extensions built on algorithm adaptation methods or problem transformation methods.

The ensemble of classifier chains (ECC) [32] is an extension that has classifier chains (CC) as base method. The driving principle is to have for each independent CC, a random chain ordering and a training on a random selection of the training set. A label is determined relevant if it was predicted by a percentage of classifiers according to a threshold value.

Random forest predictive clustering tree (RF-PCT) [28, 37] is an extension that has predictive clustering tree (PCT) as base method. PCT is a decision tree organized as a hierarchy of clusters. In a random forest, each tree in the ensemble is learned using a bootstrap sample from the training set. Each PCT provides a multi-label classification and then the fusion of their result is done by using some voting scheme.

Random k-Labelsets (RAkEL) [38] is an extension that is based on LP. RAkEL tackles the computational efficiency issue of LP by breaking the original set of labels into k smaller random subsets where LP can be applied

efficiently. A simple voting process is used to determine the final set of relevant labels.

3. Materials

In this section, we describe the characteristics of the MIMIC-II clinical database [24]. We also explain how we use these data for our study related to chronic diseases.

The data were gathered during a seven year period, beginning in 2001, from Intensive Care Unit (ICU) of Boston's Beth Israel Deaconess Medical Center (BIDMC). The MIMIC-II clinical database [24] is publicly and freely available after registration. The last release of the database contains around 33,000 patients. We choose to skip the neonates and the children in order to concentrate only on the adult population ($\geq$ 16 years old) which consists of around 24,000 patients, where we extracted a subset of 19,773 patients with chronic diseases. Regarding the restriction to the adult population, we motivate this decision by the divergence which exists between these two groups in term of medical conditions and treatment plans. The average age of the patients in the database is 67 years old. A high proportion of elderly patients in ICU can be explained by the worsening of their comorbidities [39]. The distribution of the population is: 56% of men and 44% of women.

The clinical data we consider are the laboratory tests and the items registered in the chart. By chart, we mean a logbook per patient which records the results of heterogeneous examinations, such as: fluid assessment, physiological measure, or severity score which evaluates vital functions. Important information such as the age and the gender of the patient is part of the chart.

According to the length of the stay, a patient will make several laboratory tests and various examinations. Thus, clinical data of patients are time series. In order to attenuate the amount of missing values, we take a subset of items, from the laboratory tests and from the chart, that are present at least for 80% of the patients. We end up with 76 items from the laboratory test and from the chart. A detailed table, with the descriptive statistics, is available in Appendix A. Note that we do not apply any feature selection algorithms, since these algorithms will select a subset of features that will optimize the accuracy for this particular database. Instead, we want to keep this work as general as possible to allow generalization to other clinical databases, in particular ICU databases.

Note that the data is not missing at random according to the documentation of the database[3]. Thus techniques such as interpolation or imputation will not give appropriate results. Several approaches exist for handling the problem of missing values in medical datasets [40]. A trivial approach is to substitute the mean for the missing values [41], however this is rarely an acceptable solution [42]. A better approach is to look at medical knowledge to substitute with values within a plausible range [42]. Given these considerations, we consider a plausible range of physiological values in case of missing data according to information we gathered in the medical literature. We either impute physiological values in ranges that are plausible given the patient disease or we impute physiological values of a healthy person if appropriate.

As labels we consider 10 families of chronic diseases where their distri-

[3]http://mimic.physionet.org/UserGuide/node16.html

butions amongst the 19,773 extracted patients are presented in the Table 1. We use the coding scheme of the International Classification of Disease revision 9 (ICD-9)[4] available in the MIMIC-II database for building the 10 families of chronic diseases as described in the Table 2. Providing a definitive diagnostic is not realistic in our settings. The ICD-9 codes system, which is especially used for morbidity statistics or health insurance systems, is not sophisticated enough to define a label for a precise diagnostic. In addition, even when a diagnostic is available, for a given disease the treatment is often specific for each patient, due to different symptoms, results of other examinations, interaction of medications, or allergies. Based on these considerations, we decided to form 10 families of chronic diseases by taking into account the medical relevance, the characteristics of the dataset and the hierarchical structure of the ICD-9 coding system. Providing information on belonging of a patient to one or several disease families will support the physician by suggesting the directions for further investigations.

Related to the definition of labels, it is important to compute the *label cardinality* and the *label density*. Label cardinality quantifies the number of alternative labels that characterize the observations in the dataset on average. With respect to label cardinality, label density considers also the number of labels. The two measures are useful because multi-label algorithms may present a different behavior in datasets with similar cardinality, but different density.

Label Cardinality: is the average number of labels of the observations in

[4]http://www.who.int/classifications/icd

Label / Chronic disease	**No. of patients**	**%**
Hypertensive disease	12,309	62.3%
Fluid electrolyte disease	6,177	31.2%
Diabetes mellitus	6,056	30.6%
Lipoid metabolism disease	5,965	30.2%
Kidney disease	5,828	29.5%
COPD	4,253	21.5%
Thyroid disease	2,246	11.4%
Hypotension	1,962	9.9%
Liver disease	1,088	5.5%
Thrombosis	931	4.7%

Table 1: Distribution of labels / families of chronic diseases in the 19,773 extracted patients of the MIMIC-II database.

a dataset D:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i|. \tag{7}$$

Label Density: is the average number of labels of the observations in a dataset D divided by $|L|$:

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|}. \tag{8}$$

The dataset presents a label cardinality of 2.37 and a label density of 0.237, with 1023 possible combinations, of which 522 are present in the dataset.

Label / Chronic disease	**ICD-9 codes**
Hypertensive disease	[401 to 405]
Fluid electrolyte disease	276
Diabetes mellitus	[249 to 250]
Lipoid metabolism disease	272
Kidney disease	[580 to 589]
COPD	[490 to 496]
Thyroid disease	[240 to 246]
Hypotension	458
Liver disease	571
Thrombosis	[451 to 453]

Table 2: ICD-9 codes considered for building the 10 families of chronic diseases.

4. Methods

In this section we describe the feature extraction and the standardization that we apply on the data, then we describe the multi-label learning algorithms considered in this study.

4.1. Feature extraction

Laboratory events and chart events of each patient are summarized into one feature vector. Due to the heterogeneity and the different frequencies of the selected medical data, we propose the following approach for the feature extraction according to the type of the measured values:

Numerical variables consist of measured values such as blood pressure, creatinine and temperature. When they appear one time, such as the height

at the patient admission, they are taken in the feature vector as they are. When they appear several times, the following summary features are computed: mean, median, standard deviation and range (max−min).

Categorical variables consist of observed values such as cardiovascular function assessment score and urine color. For a patient which did several times a particular examination where results are discrete values which can be divided into mutually exclusive classes, we can represent this information as an histogram. Then, the relative frequency of each category of the histogram is used as feature. There is also the case where only one observation exists for each patient, such as the gender at the patient admission, in that case, we encode as feature the value in a binary variable.

4.2. Standardization

Distance based algorithms, such as ML-kNN, may be affected by distortions on distances due to the scale of features can be different. Then, we compute the z-score for each element of our feature vectors set, as defined below.

Let a data set $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{x}_i$ is a d-dimensional feature vector $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{id}]$. The z-score value z_{ij} of a value x_{ij} is:

$$z_{ij} = \frac{x_{ij} - \bar{X}_j}{S_j}, \tag{9}$$

where $\bar{X}_j$ is the j-th dimension sample mean and S_j is the j-th dimension sample standard deviation.

4.3. Multi-label classifiers

According to the analysis performed in [11], we will consider the following multi-label learning algorithms for our evaluations on clinical data:

Binary Relevance (BR) [31], Classifier Chain (CC) [32] and HOMER [34] for problem transformation methods; and Random k-labelsets (RAkEL) [38] for ensemble methods. Regarding problem adaptation methods, we will take into consideration the AdaBoostMH [26] which is the multi-label adaptation of AdaBoost. We will also consider the ML-kNN [21] which is the multi-label adaptation of kNN.

For transformation methods which need a base classifier, we decided to use Support Vector Machines (SVM) [43], Naive Bayes (NB) [44] and Decision Trees (J48) [45]. The choice of SVM with an RBF kernel is motivated by its ability to model non-linear problems where classes are not linearly separable. SVM needs two parameters, the first one is the coefficient of the RBF kernel (γ), the second one is the penalty of the error term (c). The choice of NB is justified by the fact the algorithm works well in many scenarios due to its simple structure and its surprising classification performance, even with the conditional independence assumption [44]. Similarly to NB, decision trees (using J48 implementation) are simple algorithms that scale optimally to large dataset, and furthermore they provide an easy explanation of the rules used for the classification. J48 needs two parameters, the first one is the confidence threshold for pruning (p), and the second one is the minimum number of instances per leaf (l).

Concerning transformation methods, no additional parameters are needed for BR and CC, HOMER needs a parameter for defining the number of clusters (k) for the k-means algorithm [46] which is used during the initialization stage of HOMER. Regarding ensemble methods, RAkEL needs two parameters where the first one defines the number of models you want to consider in

the ensemble, and the second one defines the size of the subset of labels considered inside each specific model of the ensemble. According to Tsoumakas et al. [38], a reasonable choice, which provides a trade-off between the computational complexity and the predictive performance, is to specify the number of models as:

$$min(2 * |L|, 100), \tag{10}$$

where L is the label set, and to specify the size of the subset of labels as:

$$|L|/2, \tag{11}$$

where L is the label set. In our case, as $|L| = 10$, the number of models is 20 and the size of the subset of labels is 5.

Concerning adaptation methods, AdaBoostMH does not need an additional parameter as it optimizes the hamming loss of decision stumps (Weka implementation[5]). ML-kNN needs the number of neighbors (N) and a smooth parameter (σ) which controls the strength of the uniform prior (Laplace approximation of the prior). ML-kNN is a multi-label algorithm which is widely used and will serve as baseline for our evaluation. Figure 1 represents how these selected learning algorithms are divided into groups using the categorization presented in the background section.

5. Experiments

In this section we describe how the experiments were conducted and we discuss about the results.

[5]http://www.cs.waikato.ac.nz/ml/weka/

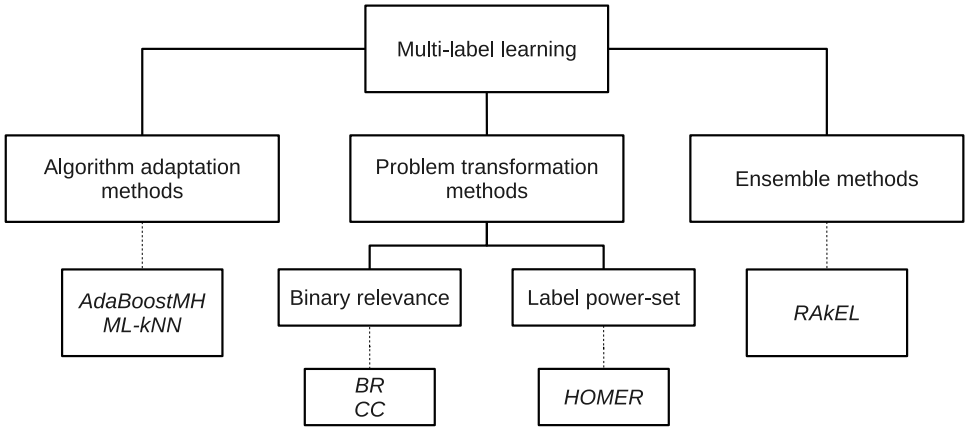

Figure 1: The multi-label learning algorithms divided into groups using the categorization presented in the background section.

Regarding the software environment in use, all the multi-label learning algorithms and evaluation metrics have been implemented with the Java programming language. The following Java libraries have been used: Mulan[6] (version 1.4) and Weka[7] (version 3.7.6). The operating system is a Ubuntu Linux 12.04 LTS 64 bits. Regarding the hardware environment, we used a workstation equipped with a Intel Core i7 CPU 870 at 2.93 GHz and 16 GB of memory (RAM). Concerning the training time, as shown in the Table 6, although the CPU has 4 cores (8 threads due to hyper-threading), only 1 core is used in practice by the fact that the Java implementation of the algorithms is single thread.

[6]http://mulan.sourceforge.net

[7]http://www.cs.waikato.ac.nz/ml/weka/

5.1. Parameters selection

The large size of the MIMIC-II dataset allows us to divide the dataset randomly in three parts where each contains enough observations for fitting optimally a model. The first part, called training set, is used to build all models across all parameters. The model selection is done using a grid search over a defined parameter space, described in Table 3. The second part, called validation set, is used to select the best parameters for each algorithm. Finally the third part, called test set, is used for computing the reported results according to the best parameters. Table 3 shows the parameters selected by the grid search for each of the algorithms. This process, which consists of the permutations of the three sets, is repeated 6 times. The reported results, in the Tables 4 and 5, are the mean of the six iterations under a confidence interval of 95%. The process schema of the experiment is presented at the Figure 2.

5.2. Results

Amongst the evaluation metrics that we consider, the hamming loss is for us the most important one because it represents the ability of the algorithm to discriminate the symbols associated to the illnesses that the patient has. The second one is the ranking loss as it concerns the ranking of the labels according to the dominant disease of the patient.

By looking at the results, in the Tables 4 and 5, we can say that decision trees perform optimally considering all the evaluation metrics, and they have the following advantages: they scale well to large datasets and they are easy to interpret. SVM-based approaches give the best accuracy for the hamming loss but they do not rank well the illnesses, and in addition, they take a very

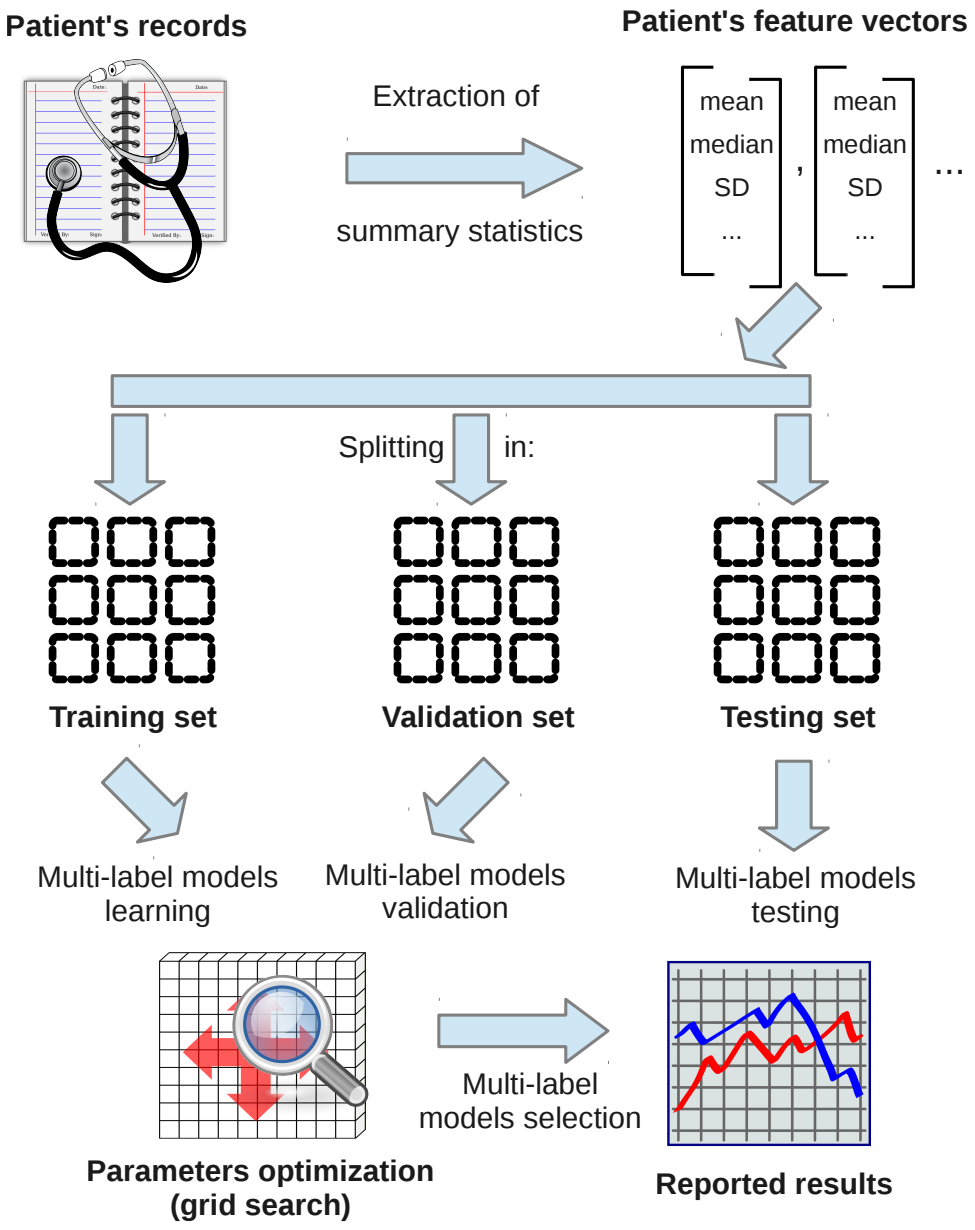

Figure 2: The overall process of the experiment.

Algorithm	**Parameters**	**Interval**	**Step**
AdaBoostMH	N/A	N/A	N/A
BR-NB	N/A	N/A	N/A
CC-NB	N/A	N/A	N/A
RAkEL-NB	N/A	N/A	N/A
HOMER-NB	$k = 2$	[2:9]	lin
ML-kNN	$N = 37$, σ=1	[1:50], [1:10]	lin, lin
BR-J48	$p = 0.001$, $l = 34$	[0.0001:0.1], [1:50]	log, lin
BR-SVM	$\gamma = 0.01$, $c = 15$	[0.001:10], [1:50]	log, lin
CC-J48	$p = 0.001$, $l = 34$	[0.0001:0.1], [1:50]	log, lin
CC-SVM	$\gamma = 0.01$, $c = 15$	[0.001:10], [1:50]	log, lin
RAkEL-J48	$p = 0.001$, $l = 34$	[0.0001:0.1], [1:50]	log, lin
RAkEL-SVM	$\gamma = 0.01$, $c = 15$	[0.001:10], [1:50]	log, lin
HOMER-J48	$p = 0.001$, $l = 34$, $k = 8$	[0.0001:0.1], [1:50], [2:9]	log, lin, lin
HOMER-SVM	$\gamma = 0.01$, $c = 15$, $k = 8$	[0.001:10], [1:50], [2:9]	log, lin, lin

Table 3: Selected parameters after the grid search.

long time to train when the Gramian matrix is big, as in the case of the MIMIC-II dataset, as shown in the Table 6.

Concerning the ML-kNN, often used as the gold standard in multi-label classification tasks, it obtains a performance very close to decision trees, but it is not very scalable, as the other instance-based learning approaches. AdaBoostMH is not competitive in these settings, we think that this is caused by the fact that this algorithm is optimizing mainly the hamming loss rather than the ranking loss, as discussed in [26]. Regarding naive Bayes

Algorithm/Metric	**Hamming Loss**	**Ranking Loss**	**Average Precision**
BR-SVM	**16.94% ± 0.12%**	34.47% ± 0.64%	61.85% ± 0.56%
HOMER-SVM	**16.97% ± 0.11%**	34.18% ± 0.61%	62.01% ± 0.56%
CC-SVM	**17.01% ± 0.14%**	33.67% ± 0.70%	62.58% ± 0.59%
RAkEL-SVM	17.18% ± 0.07%	27.08% ± 0.45%	68.32% ± 0.54%
BR-J48	17.63% ± 0.13%	**17.70% ± 0.31%**	**72.23% ± 0.55%**
HOMER-J48	17.75% ± 0.17%	33.06% ± 1.05%	62.45% ± 0.89%
CC-J48	17.83% ± 0.15%	**17.76% ± 0.40%**	**72.14% ± 0.79%**
RAkEL-J48	18.17% ± 0.09%	25.88% ± 1.32%	68.53% ± 0.99%
ML-kNN	18.91% ± 0.16%	**18.10% ± 0.17%**	71.37% ± 0.22%
AdaBoostMH	21.23% ± 0.09%	32.81% ± 0.13%	57.65% ± 0.19%
HOMER-NB	21.78% ± 0.30%	34.76% ± 0.47%	57.61% ± 0.99%
RAkEL-NB	24.88% ± 0.40%	25.63% ± 0.39%	64.07% ± 0.76%
BR-NB	28.40% ± 0.34%	27.62% ± 0.50%	60.32% ± 0.70%
CC-NB	28.61% ± 0.36%	27.92% ± 0.51%	59.96% ± 0.71%
Random	*49.89% ± 0.19%*	*49.99% ± 0.43%*	*40.13% ± 0.40%*

Table 4: Results for the hamming loss, the ranking loss and the average precision.

approaches, the non-competitive results may be explained by the fact that the NB assumes statistical independence between the features. However, in our dataset, we know that some features are highly correlated according to medical observations such as the very well known link between diabetes and hypertension, as described in literature [47]. We also evaluate the system against a random multi-label classification in order to know the floor or the ceiling which must not be excessed.

Algorithm/Metric	**One-Error**	**Coverage**
RAkEL-SVM	**30.68% ± 0.81%**	4.72 ± 0.04
BR-J48	**31.15% ± 0.70%**	**3.50 ± 0.01**
CC-J48	**31.48% ± 1.32%**	**3.50 ± 0.02**
ML-kNN	32.05% ± 0.36%	**3.52 ± 0.02**
RAkEL-J48	32.28% ± 0.66%	4.57 ± 0.13
HOMER-J48	34.74% ± 0.87%	5.31 ± 0.10
CC-SVM	35.11% ± 0.80%	5.28 ± 0.06
BR-SVM	35.17% ± 0.67%	5.35 ± 0.05
HOMER-SVM	35.34% ± 0.74%	5.33 ± 0.05
AdaBoostMH	37.75% ± 0.29%	4.89 ± 0.01
HOMER-NB	44.35% ± 2.89%	5.29 ± 0.03
RAkEL-NB	45.93% ± 2.19%	4.37 ± 0.03
BR-NB	54.80% ± 1.37%	4.42 ± 0.03
CC-NB	55.55% ± 1.40%	4.44 ± 0.03
Random	*76.03% ± 0.54%*	*6.30 ± 0.03*

Table 5: Results for the one-error and the coverage.

Regarding the transformation methods, CC, which is an extension to consider the correlation between the labels, as described in the section 2, is not able to improve significantly the results compared to the BR method, which considers each label independently. As discussed in [48], an explanation could be a suboptimal ordering of the labels in the chain, another reason could be an accumulation of errors induced by the chain of CC which tends to increase when the number of labels is large. In addition, after extensive benchmarks

Algorithm	**Training time**
BR-NB	1 min
CC-NB	1 min
HOMER-NB	1 min
BR-J48	4 min
CC-J48	4 min
HOMER-J48	4 min
AdaBoostMH	6 min
RAkEL-NB	7 min
RAkEL-J48	15 min
ML-kNN	35 min
BR-SVM	3 h 32 min
CC-SVM	3 h 33 min
HOMER-SVM	4 h 11 min
RAkEL-SVM	28 h 13 min

Table 6: Training time of each algorithm according to the best parameters given by the grid search.

on various datasets, the authors of [48] concluded that the performance of CC dramatically drops when the complexity of the dataset increases, such as: a larger number of labels, a greater cardinality or a higher label dependency. The hierarchy of multi-label classifiers (HOMER) is not bringing a significant improvement with respect to other approaches, except when NB is used as the base classifier. We think that for the way HOMER works, the creation of artificial meta labels as parents of smaller groups of related labels attenuates the previously discussed issue of NB.

The RAkEL algorithm, being an ensemble method, suffers from scalability problems as we have a large dataset. RAkEL extends the LP method by breaking the original set of labels in several random subsets where LP is used internally. Although RAkEL improves substantially over LP, the drawbacks inherent of LP remain. Moreover RAkEL needs two additional parameters which have been defined as fixed values according to the recommendation of RAkEL's authors [38], since in the case of our study we have a large dataset. In particular, according to the Table 6, a grid search over 4 parameters (instead of 2) with an SVM as the base classifier would have been impracticable under our available computing power. However, despite the scalability issue, RAkEL with an SVM as the base classifier obtains competitive results to the BR methods, in particular it improves the results for the average precision and the ranking loss, which is relevant for the ranking of the labels according to the dominant disease of the patient. In addition, RAkEL with an SVM as the base classifier achieves the best performance, over all other methods considered in our study, for the one-error evaluation metric which evaluates the reliability of the top-ranked label.

5.3. Discussion

Our recommendation regarding the analysis of these results concerning medical data and chronic diseases is to choose decision tree based algorithms, because they performed well accross all evaluation metrics and scale up to large dataset. However, if the scenario of the hamming loss is more important, when we want to achieve the best performance regarding the identification of all diseases a patient may be affected, regardless their rankings or their significance levels, the best algorithms are those based on SVM with

BR, but with the inconvenience to be slow at training time. Another consideration to make is that in the case of chronic diseases, like in MIMIC-II, pure multi-label algorithms such as ML-kNN or AdaBoostMH do not seem to have an advantage to the BR approaches. On the other hand, the bad performance of the NB algorithm gives us a direction: in multi-label medical domains, the correlation between the features is an important characteristic to take into consideration.

Quite surprisingly we discovered that the most used multi-label learning algorithms, with the exception of RAkEL, do not improve the results with the respect of a binary relevance approach that makes the independence assumption of the chronic illnesses. It is difficult to give a complete explanation about these results, however we can provide two opposite reasons for this behavior. The first one could be that the feature extraction process cannot model optimally the correlation between the features in order that multi-label learning approaches can exploit this information. The second one could be that the extracted features are sufficient for discriminating the various illnesses. To confirm this, we think that further studies are required with algorithms and processing methods that can deepen the analysis of dependencies of physiological measurements and chronic diseases. Furthermore, in view of the promising results of the RAkEL algorithm for the ranking of diseases, we think that further developments in the direction of algorithms exploiting ensemble methods and label power-set methods should be considered.

We experienced that, in the particular context of large medical datasets, it is more convenient to use algorithms with few parameters. By the use

of sequential data, in order to maintain their interpretability, we decided to use a summary statistics approach for the feature extraction rather than an unsupervised approach such as vector quantization techniques. We think that for this dataset, the conception of a well trained binary relevance method is sufficient to obtain a decent model. More sophisticated multi-label learning methods, such as the promising RAkEL method, can bring an added value but at the cost of a greater complexity.

6. Conclusion

In this contribution we presented an evaluation of multi-label learning algorithms on patients affected by chronic diseases. The emphasis of the work is on trying to model the relationship between different chronic illnesses by means of the multi-label paradigm. In this study we have been faced with the MIMIC-II dataset which contains a large number of patient records. This aspect leads to scalability problems with classifiers where multiple parameters need to be optimized, such as the use of a grid search method, to obtain an optimal model.

Future work implies attempting algorithms in the direction of: combining the advantages of the BR and the LP methods, exploiting better the correlation between the features and restraining the number of parameters to optimize. From these considerations, we proposed in [49] a preliminary attempt of a probabilistic multi-label learning framework for the analysis of medical data. In addition to trying multi-label algorithms, another interesting direction is to infer a correlated illness to the known one without having access to the key feature that can discriminate it. A further work would be

to consider the multi-label analysis of continuous non-invasive signals such as ECG, breathing and activity to see if multi-label methods can help to diagnose comorbid pathological conditions. More precisely, working with ECG signals from the MIMIC-II Waveform Database, will allow to detect the presence (or co-presence) of cardiovascular diseases such myocardial infarction, coronary artery disease, or arrhythmia.

Acknowledgment

This work was partially supported by the EU FP7 287841 COMMODITY12 project.

Appendix A. Dataset descriptive statistics

In this appendix, we present the list of variables from the MIMIC-II dataset that are considered in our study. The descriptive statistics presented below are computed across all measured values for all patients. In the table A.7, the descriptive statistics (mean, median and standard deviation) for the quantitative values are given. In the table A.8, the descriptive statistics (frequency) for the categorical values are given.

Measured item	**Mean**	**Median**	**SD**
Age at the admission (years)	67.48	68.0	22.35
Height at the admission (inches)	66.70	67.0	7.90
Weight at the admission (kg)	81.58	78.4	24.08
Body surface area at the admission (m^2)	1.99	2.0	0.32
Heart rate (BPM)	86.36	85.0	17.87
Blood Pressure Systolic (mmHg)	120.03	118.0	23.82

Blood Pressure Diastolic (mmHg)	57.77	56.0	15.64
Respiratory Rate (BPM)	20.14	20.0	6.39
Saturation of peripheral oxygen (%)	97.20	98.0	3.74
Temperature (deg. C)	36.96	36.9	1.1
Hematocrit [Volume Fraction] of Blood (%)	31.37	30.8	5.12
Platelets [#/volume] in Blood (K/uL)	237.01	214.0	145.17
Leukocytes [#/volume] in Blood (K/uL)	10.34	8.9	20.88
Hemoglobin [Mass/volume] in Blood (g/dL)	10.61	10.4	1.78
Erythrocyte mean corpuscular volume [Entitic volume] (fL)	90.04	90.0	6.73
Erythrocytes [#/volume] in Blood (m/uL)	3.54	3.5	0.64
Erythrocyte mean corpuscular hemoglobin concentration [Mass/volume] (%)	33.51	33.6	1.56
Erythrocyte mean corpuscular hemoglobin [Entitic mass] (pg)	30.15	30.2	2.5
Erythrocyte distribution width [Ratio] (%)	15.84	15.4	2.34
Urea nitrogen [Mass/volume] in Serum or Plasma (mg/dL)	30.69	23.0	23.38
Creatinine [Mass/volume] in Serum or Plasma (mg/dL)	1.67	1.1	1.78
Potassium [Moles/volume] in Serum or Plasma (mEq/L)	4.17	4.1	0.67
Sodium [Moles/volume] in Serum or Plasma (mEq/L)	138.6	139.0	4.85
Chloride [Moles/volume] in Blood (mEq/L)	103.36	103.0	6.02

Bicarbonate [Moles/volume] in Serum (mEq/L)	25.37	25.0	4.9
Anion gap in Blood (mEq/L)	14.1	14.0	3.81
Glucose [Mass/volume] in Serum or Plasma (mg/dL)	131.41	116.0	66.81
Magnesium [Mass/volume] in Serum or Plasma (mg/dL)	2.01	2.0	0.39
INR in Blood by Coagulation assay	1.72	1.3	1.86
Prothrombin time (PT) in Blood by Coagulation assay (seconds)	16.69	14.5	7.13
Activated partial thrombplastin time (aPTT) in Blood by Coagulation assay (seconds)	45.17	34.6	26.98
Phosphate [Mass/volume] in Serum or Plasma (mg/dL)	3.66	3.4	1.35
Calcium [Mass/volume] in Serum or Plasma (mg/dL)	8.54	8.5	0.85
pH of Urine (units)	5.91	5.5	1.0
Urobilinogen [Mass/volume] in Urine (mg/dL)	0.4	0.0	1.45
Ketones [Mass/volume] in Urine (mg/dL)	3.84	0.0	18.96
Specific gravity of Urine by Test strip	1.02	1.0	0.12
Protein [Mass/volume] in Urine by Test strip (mg/dL)	32.6	0.0	93.87
Glucose [Mass/volume] in Urine (mg/dL)	54.25	0.0	205.37

Table A.7: MIMIC-II dataset descriptive statistics for quantitative values.

Measured item	**Frequency**
Gender	Male = 55.89%, Female = 44.11%
Marital status	Single/Divorced/Widowed = 51.49%, Married = 48.51%
Heart rhythm	Normal sinus = 59.94%, Abnormal sinus (such as arrhythmia) = 40.06%
Ectopic heartbeat	No = 84.32%, Yes (PAC, PNC, PVC) = 15.68%
Level of conscious	Alert = 46.64%, Arouse to pain/stimuli/voice = 23.13%, Unresponsive = 30.23%
Eye opening (Glasgow coma scale [50])	Spontaneously = 62.20%, To speech/pain = 25.09%, No response = 12.71%
Verbal response (Glasgow coma scale [50])	Oriented = 40.67%, Confused/Inappropriate = 8.81%, No response = 50.52%
Motor response (Glasgow coma scale [50])	Obeys commands = 68.70%, Localizes pain/Flex-withdraws = 23.51%, No response = 7.79%
Respiratory pattern	Regular = 94.37%, Irregular = 5.63%
Cardiovascular SOFA score [51] (low is better)	0 = 26.64%, 1 or 2 = 73.36%, 3 or 4 = 0.00%
Hematologic SOFA score [51] (low is better)	0 = 68.85%, 1 or 2 = 27.13%, 3 or 4 = 4.02%
Neurological SOFA score [51] (low is better)	0 = 31.80%, 1 or 2 = 32.05%, 3 or 4 = 36.15%
Renal SOFA score [51] (low is better)	0 = 55.68%, 1 or 2 = 26.66%, 3 or 4 = 17.66%

LUL lung sounds	Clear = 58.19%, Not clear = 41.81%
LLL lung sounds	Clear = 24.76%, Not clear = 75.24%
RUL lung sounds	Clear = 58.28%, Not clear = 41.72%
RLL lung sounds	Clear = 24.76%, Not clear = 75.24%
Skin integrity	Intact = 55.52%, Impaired = 44.48%
Bowel sounds	Present = 91.66%, Absent = 8.34%
Activity (Braden scale [52])	Walks frequently/occasionally = 5.60%, Chairfast/Bedfast = 94.40%
Moisture (Braden scale [52])	Rarely moist = 44.55%, Occasionally moist = 48.09%, Very moist = 7.36%
Mobility (Braden scale [52])	No limitation = 4.30%, Slightly limited = 35.96%, Very limited = 59.74%
Sensory perception (Braden scale [52])	No impairment = 21.71%, Slightly limited = 39.33%, Very limited = 38.96%
Nutrition (Braden scale [52])	Excellent/Adequate = 33.53%, Probably inadequate = 57.78%, Very poor = 8.69%
Friction and shear (Braden scale [52])	No apparent problem = 16.35%, Potential problem = 70.74%, Problem = 12.91%
Assistance device	Independent = 5.32%, Supervised/Assisted = 94.68%
Urine color	(Light) yellow = 79.16%, Not yellow = 20.84%
Urine appear	Clear = 85.25%, Sediment/Cloudy = 14.75%
Intravenous site appear	Within normal range = 96.52%, Outside normal range = 3.48%
Pain present	No = 74.06%, Yes = 25.94%

Contact precautions	No = 76.48%, Yes = 23.52%
Sedation-agitation (Riker scale [53])	Calme/Cooperative = 68.15%, Agitated = 6.51%, Sedated = 25.34%
Restraint location	No = 35.53%, Yes = 64.47%
Nitrite [Presence] in Urine by Test strip	Negative = 94.29%, Positive = 5.71%
Bilirubin [Presence] in Urine	Negative = 89.74%, Positive = 10.26%
Hemoglobin [Presence] in Urine by Test strip	Negative = 43.16%, Positive = 56.84%
Leukocytes [Presence] in Urine	Negative = 70.69%, Positive = 29.31%

Table A.8: MIMIC-II dataset descriptive statistics for categorical values.

References

[1] S. S. Lim, T. Vos, A. D. Flaxman, G. Danaei, K. Shibuya, H. Adair-Rohani, M. A. AlMazroa, M. Amann, H. R. Anderson, K. G. Andrews, et al., A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010, The Lancet 380 (9859) (2013) 2224 – 2260. doi:10.1016/S0140-6736(12)61766-8.

[2] C. Roehrig, G. Miller, C. Lake, J. Bryant, National health spending by

medical condition, 1996-2005, Health Affairs 28 (2) (2009) w358–w367. doi:10.1377/hlthaff.28.2.w358.

[3] C. Mathers, D. M. Fat, J. Boerma, The global burden of disease: 2004 update, World Health Organization, 2008.

[4] G. Danaei, M. M. Finucane, Y. Lu, G. M. Singh, M. J. Cowan, C. J. Paciorek, J. K. Lin, F. Farzadfar, Y.-H. Khang, G. A. Stevens, M. Rao, M. K. Ali, L. M. Riley, C. A. Robinson, M. Ezzati, National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants, The Lancet 378 (9785) (2011) 31 – 40. doi:10.1016/S0140-6736(11)60679-X.

[5] A. Alwan, et al., Global status report on noncommunicable diseases 2010., World Health Organization, 2011.

[6] K. G. M. M. Alberti, et al., Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus., World Health Organization, 1999.

[7] Ö. Kafalı, S. Bromuri, M. Sindlar, T. van der Weide, E. Aguilar Pelaez, U. Schaechtle, B. Alves, D. Zufferey, E. Rodriguez-Villegas, M. I. Schumacher, K. Stathis, Commodity 12: A smart e-health environment for diabetes management, Journal of Ambient Intelligence and Smart Environments 5 (5) (2013) 479–502. doi:10.3233/AIS-130220.

[8] J. Ghably, B. Paterson, A. Peiris, Endocrinology in crisis?, Southern medical journal 106 (4) (2013) 245.

[9] T. G. Kannampallil, A. Franklin, R. Mishra, K. F. Almoosa, T. Cohen, V. L. Patel, Understanding the nature of information seeking behavior in critical care: Implications for the design of health information technology, Artificial Intelligence in Medicine 57 (1) (2013) 21 – 29. doi:10.1016/j.artmed.2012.10.002.

[10] G. Tsoumakas, I. Katakis, I. Vlahavas, Mining multi-label data, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer US, 2010, pp. 667–685.

[11] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Džeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognition 45 (9) (2012) 3084 – 3104, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011). doi:10.1016/j.patcog.2012.03.004.

[12] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, Knowledge and Data Engineering, IEEE Transactions on 26 (8) (2014) 1819–1837. doi:10.1109/TKDE.2013.39.

[13] J. Wang, P. Liu, M. F. She, S. Nahavandi, A. Kouzani, Bag-of-words representation for biomedical time series classification, Biomedical Signal Processing and Control 8 (6) (2013) 634 – 644. doi:10.1016/j.bspc.2013.06.004.

[14] J. Sun, F. Wang, J. Hu, S. Edabollahi, Supervised patient similarity measure of heterogeneous patient records, SIGKDD Explor. Newsl. 14 (1) (2012) 16–24. doi:10.1145/2408736.2408740.

[15] A. Wieczorkowska, P. Synak, Z. Raś, Multi-label classification of emotions in music, in: M. Kłopotek, S. Wierzchoń, K. Trojanowski (Eds.), Intelligent Information Processing and Web Mining, Vol. 35 of Advances in Soft Computing, Springer Berlin Heidelberg, 2006, pp. 307–315.

[16] K. Trohidis, G. Tsoumakas, G. Kalliris, I. P. Vlahavas, Multi-label classification of music into emotions, in: ISMIR, Vol. 8, 2008, pp. 325–330.

[17] M. R. Boutell, J. Luo, X. Shen, C. M. Brown, Learning multi-label scene classification, Pattern Recognition 37 (9) (2004) 1757 – 1771. doi:10.1016/j.patcog.2004.03.009.

[18] A. McCallum, Multi-label text classification with a mixture model trained by EM, in: AAAI'99 Workshop on Text Learning, 1999, pp. 1–7.

[19] Z. Barutcuoglu, R. E. Schapire, O. G. Troyanskaya, Hierarchical multi-label prediction of gene function, Bioinformatics 22 (7) (2006) 830–836. doi:10.1093/bioinformatics/btk048.

[20] X. Xiao, Z.-C. Wu, K.-C. Chou, iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, Journal of Theoretical Biology 284 (1) (2011) 42 – 51. doi:10.1016/j.jtbi.2011.06.005.

[21] M.-L. Zhang, Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition 40 (7) (2007) 2038 – 2048. doi:10.1016/j.patcog.2006.12.019.

[22] M.-J. Huang, M.-Y. Chen, S.-C. Lee, Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis, Expert Systems with Applications 32 (3) (2007) 856 – 867. doi:10.1016/j.eswa.2006.01.038.

[23] J. L. Amaral, A. J. Lopes, J. M. Jansen, A. C. Faria, P. L. Melo, Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease, Computer Methods and Programs in Biomedicine 105 (3) (2012) 183 – 193. doi:10.1016/j.cmpb.2011.09.009.

[24] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, R. G. Mark, Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database, Critical Care Medicine 39 (2011) 952–960.

[25] S. Bromuri, D. Zufferey, J. Hennebert, M. Schumacher, Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms, Journal of Biomedical Informatics 51 (0) (2014) 165 – 175. doi:10.1016/j.jbi.2014.05.010.

[26] R. Schapire, Y. Singer, Boostexter: A boosting-based system for text categorization, Machine Learning 39 (2-3) (2000) 135–168. doi:10.1023/A:1007649029923.

[27] A. Clare, R. King, Knowledge discovery in multi-label phenotype data, in: L. De Raedt, A. Siebes (Eds.), Principles of Data Mining and

Knowledge Discovery, Vol. 2168 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2001, pp. 42–53.

[28] H. Blockeel, L. D. Raedt, J. Ramon, Top-down induction of clustering trees, in: Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 55–63.

[29] M.-L. Zhang, Z.-H. Zhou, Multilabel neural networks with applications to functional genomics and text categorization, Knowledge and Data Engineering, IEEE Transactions on 18 (10) (2006) 1338–1351. doi:10.1109/TKDE.2006.162.

[30] A. Elisseeff, J. Weston, A kernel method for multi-labelled classification, in: Advances in Neural Information Processing Systems 14, MIT Press, 2001, pp. 681–687.

[31] G. Tsoumakas, I. Katakis, Multi-label classification: An overview, International Journal of Data Warehousing and Mining (IJDWM) 3 (3) (2007) 1–13. doi:10.4018/jdwm.2007070101.

[32] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Machine Learning 85 (3) (2011) 333–359. doi:10.1007/s10994-011-5256-5.

[33] J. Read, A pruned problem transformation method for multi-label classification, in: Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008), 2008, pp. 143–150.

[34] 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 (MMD08), 2008, pp. 30–44.

[35] E. Loza Mencía, J. Fürnkranz, Efficient pairwise multilabel classification for large-scale problems in the legal domain, in: W. Daelemans, B. Goethals, K. Morik (Eds.), Machine Learning and Knowledge Discovery in Databases, Vol. 5212 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2008, pp. 50–65. doi:10.1007/978-3-540-87481-2_4.

[36] E. L. Mencía, S.-H. Park, J. Fürnkranz, Efficient voting prediction for pairwise multilabel classification, Neurocomputing 73 (79) (2010) 1164 – 1176. doi:10.1016/j.neucom.2009.11.024.

[37] D. Kocev, C. Vens, J. Struyf, S. Džeroski, Ensembles of multi-objective decision trees, in: J. Kok, J. Koronacki, R. Mantaras, S. Matwin, D. Mladenič, A. Skowron (Eds.), Machine Learning: ECML 2007, Vol. 4701 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 624–631. doi:10.1007/978-3-540-74958-5_61.

[38] G. Tsoumakas, I. Katakis, L. Vlahavas, Random k-labelsets for multilabel classification, Knowledge and Data Engineering, IEEE Transactions on 23 (7) (2011) 1079–1089. doi:10.1109/TKDE.2010.164.

[39] L. Fuchs, C. Chronaki, S. Park, V. Novack, Y. Baumfeld, D. Scott, S. McLennan, D. Talmor, L. Celi, ICU admission characteristics and

mortality rates among elderly and very elderly patients, Intensive Care Medicine 38 (10) (2012) 1654–1661. doi:10.1007/s00134-012-2629-6.

[40] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, H. Stern, The prevention and treatment of missing data in clinical trials, New England Journal of Medicine 367 (14) (2012) 1355–1360. doi:10.1056/NEJMsr1203730.

[41] J. D. Dziura, L. A. Post, Q. Zhao, Z. Fu, P. Peduzzi, Strategies for dealing with missing data in clinical trials: From design to analysis, Yale J Biol Med. 86 (3) (2013) 343–358.

[42] D. Jackson, I. R. White, M. Leese, How much can we learn about missing data?: an exploration of a clinical trial in psychiatry, Journal of the Royal Statistical Society: Series A (Statistics in Society) 173 (3) (2010) 593–612. doi:10.1111/j.1467-985X.2009.00627.x.

[43] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/BF00994018.

[44] H. Zhang, The optimality of naive bayes, in: V. Barr, Z. Markov (Eds.), FLAIRS Conference, AAAI Press, 2004.

[45] S. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, Systems, Man and Cybernetics, IEEE Transactions on 21 (3) (1991) 660–674. doi:10.1109/21.97458.

[46] J. A. Hartigan, M. A. Wong, Algorithm as 136: A k-means clustering algorithm, Applied statistics (1979) 100–108.

[47] E. Barret-Connor, M. Criqui, M. Klauber, M. Holdbrook, Diabetes and hypertension in a community of older adults, American Journal of Epidemiology 113 (3) (1981) 276–284.

[48] O. Luaces, J. Dez, J. Barranquero, J. del Coz, A. Bahamonde, Binary relevance efficacy for multilabel classification, Progress in Artificial Intelligence 1 (4) (2012) 303–313. doi:10.1007/s13748-012-0030-x.

[49] D. Zufferey, Probabilistic multi-label learning for medical data, IEEE Intelligent Informatics Bulletin 15 (1) (2014) 26–27.

[50] G. Teasdale, G. Murray, L. Parker, B. Jennett, Adding up the glasgow coma score, in: J. Brihaye, P. Clarke, F. Loew, J. Overgaard, E. Pásztor, B. Pertuiset, K. Schürmann, L. Symon (Eds.), Proceedings of the 6th European Congress of Neurosurgery, Vol. 28 of Acta Neurochirurgica, Springer Vienna, 1979, pp. 13–16. doi:10.1007/978-3-7091-4088-8_2.

[51] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendona, H. Bruining, C. Reinhart, P. Suter, L. Thijs, The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, Intensive Care Medicine 22 (7) (1996) 707–710. doi:10.1007/BF01709751.

[52] N. Bergstrom, B. J. Braden, A. Laguzza, V. Holman, The braden scale for predicting pressure sore risk., Nursing research 36 (4) (1987) 205–210.

[53] R. R. Riker, J. T. Picard, G. L. Fraser, Prospective evaluation of the

sedation-agitation scale for adult critically ill patients, Critical care medicine 27 (7) (1999) 1325–1329.