BMC Medical Research Methodology

RESEARCH

Open Access

Single-label and multi-label classification for disease recognition with special consideration of comorbidities

Sophie Schmiegel[1], Hannah Marchi[1], Marvin-Hendrik Röchter[2,3], Martin Rudwaleit[2,3] and Christiane Fuchs[1,4*]

Abstract

Background Certain diseases require rapid treatment to avoid long-term consequences for patients. However, they may be difficult to recognize, especially if the symptoms are ambiguous and compatible with multiple possible diagnoses. Completing all necessary examinations often takes time, thereby prolonging patient suffering. Data-driven approaches, such as single-label classification (SLC) and multi-label classification (MLC), can help accelerate the diagnostic process and improve accuracy.

Methods Comparing SLC and MLC allows us to investigate whether disease recognition benefits from considering comorbidities. In this context, we aim to provide a conceptual framing of (differences between) the two approaches in model formulation, decision spaces and handling of class imbalance. To empirically assess their performance, we conduct a case study applying SLC and MLC to data from chronic pain patients.

Results Our analysis yields an ambiguous picture of whether incorporating comorbidities improves disease recognition. The suitability of SLC and MLC is determined by multiple factors, notably the dependency structure among diseases and between diseases and covariates, as well as by data characteristics such as class imbalance. Consequently, it is of high importance to address each individual problem in an individual manner, especially when the implications can be significant, such as in patient referrals.

Conclusion Our study highlights the importance of considering the specific characteristics of the data when selecting an appropriate classification approach for disease recognition and beyond. We illustrate how underlying assumptions, learning structures, and strategies for handling class imbalance shape predictive performance. Our results thus lead to the point that computational prediction tools should serve as decision support for treating physicians, but never replace their clinical judgment.

Keywords Class imbalance, Classifier chain (CC), Comorbidities, Decision space, Immune-mediated inflammatory diseases (IMID), Multi-label classification (MLC), Single-label classification (SLC)

*Correspondence:
Christiane Fuchs
christiane.fuchs@uni-bielefeld.de
[1]Faculty of Business Administration and Economics, Data Science Group, Bielefeld University, Universitätsstraße 25, Bielefeld 33615, Germany
[2]Department of Internal Medicine and Rheumatology, Hospital Bielefeld Rosenhöhe, An der Rosenhöhe 27, Bielefeld 33647, Germany
[3]Medical School and University Medical Center OWL, Bielefeld University, Universitätsstraße 25, Bielefeld 33615, Germany
[4]Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Background

Introduction

Patients can suffer from several diseases simultaneously. With regard to relevant symptoms, there is often one disease that is mainly responsible. This disease is called 'main disease' in the following. However, possible other diseases, known as 'comorbidities', might be present and cause similar symptoms. Based on this knowledge, various differential diagnoses are conceivable. Investigations to verify all suspected diagnoses may take several weeks or months, especially if waiting times for appointments at secondary care units are long. This can lead to a delay in adequate treatment prolonging the patient's suffering and potentially leading to long-term consequences. It seems reasonable to consider a patient holistically as comorbidities, assumed that they are correlated with the main disease, provide further information which can be used for the recognition of the disease of interest. This may facilitate the diagnostic procedure. Data-driven disease recognition approaches might support physicians in diagnosing a patient faster by simultaneously taking proper consideration of comorbidities.

In statistics, disease prediction is a typical case of classification (e.g. Dangare and Apte [1], Vijayarani and Dhayanand [2], Gárate-Escamila et al. [3]). Two concepts for classification are distinguished: single-label classification (SLC) and multi-label classification (MLC). The difference between these two approaches lies in the number of labels, i.e. the assignment of one unit of a dataset (henceforth sample) to a specific group (henceforth class). More specific, SLC methods assume a sample to exclusively belong to exactly one class; the respective label of this class is assigned to the sample. In contrast to that, MLC approaches are able to assign multiple labels to a sample; a sample can consequently belong to one class, to several classes simultaneously or even to none. Depending on which assumptions are fulfilled by the classification task at hand, one concept might be better suited.

In literature, SLC is used considerably more frequently than MLC. Still, both approaches are used in various domains, such as text categorization (e.g. Kwon and Lee [4], Crammer et al. [5], Liu et al. [6]), fault detection (e.g. Dineva et al. [7]), detection of emotion in music (e.g. Turnbull et al. [8], Trohidis et al. [9]) and in the medical context. As an example for MLC, Abdel Maksoud et al. [10] summarize contributions to its application to diagnose several diseases via medical image classification. Further, Farooq et al. [11] use a multi-class convolutional neural network for Alzheimer's disease prediction. Wosiak et al. [12] analyze the identification of comorbidities by applying different methods for MLC. They investigate methods that ignore relations between classes and those which account for them and demonstrate that the latter ones achieve higher values for the considered performance measures. In addition, Zufferey et al. [13] use MLC to achieve precise classification of patients with chronic diseases. They consider time series health records of patients in the MIMIC-II dataset [14].

The performance of SLC and MLC methods is compared in different research fields. To give a few examples, Sajid et al. [15] review the use of these methods for document classification as well as arising issues. They point out the need for MLC to appropriately assign categories to research articles based on metadata or the articles' content. In the context of drug development, Michielan et al. [16] recommend to apply MLC instead of SLC to predict those enzymes which are involved in the metabolism of a specific substrate. The review by Wu et al. [17] highlights the limitations of SLC as well as the advantages of MLC in the field of microbiome research, assuming that patients may suffer from several diseases simultaneously.

Despite the frequent use of SLC and MLC for disease recognition based on health records, we are not aware of a methodological review of these two approaches with respect to their conceptual foundations, practical procedures, assumed decision spaces, and strategies for handling class imbalance. Our work aims to fill this gap. Other than Zufferey et al. [13], Wosiak et al. [12] and Wu et al. [17], we explore the potential added value of considering comorbidities by applying both classification approaches and comparing their predictive performance. Consequently, the novelty and added value of our contribution lies in two complementary aspects: First, a concept-driven analysis that explains and visualizes the fundamental differences between SLC and MLC in a concise yet comprehensible manner; and second, a data-driven investigation that applies both approaches to patient-derived medical data of a hospital-based department for internal medicine and rheumatology in Bielefeld, Germany, without assuming superiority of either method in advance. By doing so, we investigate whether the consideration and simultaneous recognition of comorbidities (using MLC) improves the predictability of the disease of interest compared to a recognition of the disease in isolation (using SLC).

Case study: recognition of immune-mediated inflammatory diseases

'Immune-mediated inflammatory diseases' (IMID) is a generic term for a group of still incurable diseases with systemic (i.e. affecting a whole organ system) inflammation, for instance, rheumatoid arthritis, spondyloarthritis and others [18]. These diseases belong to a class of inflammatory autoimmune diseases which are driven by genetic and environmental factors [19] and are caused by an improper inflammatory response of the body's

immune system to self-antigens [20]. Suffering from IMID leads to pain and other symptoms, and diminishes functioning and quality of life [21]. Effective therapies which can alleviate the pain (usually consisting of basic therapeutic agents and a glucocorticoid) are already available [18].

Fast diagnosis facilitates the contemporary initiation of a proper therapy which might reduce the complaints and avoid severe long-term consequences. An important part of the standard diagnostic procedure for IMID is measuring the value of the C-reactive protein (CRP) in the blood, as this is a valid indicator for an inflammation [22]; CRP is often increased in IMID. A patient with an increased CRP and suspicious pain is usually referred from primary to secondary care unit to confirm IMID. However, long waiting times for an appointment delay the start of suitable therapy. Such waiting times arise due to limited capacity in the secondary care units, which is, among others, strained by many other, often unnecessary referrals. Inaccurate referrals, again, stem from the fact that a precise diagnosis is challenging for non-specialized primary care physicians. More reliable identification of IMID patients in primary care would relieve the secondary care units and improve the chances of providing rapid treatment for these patients.

IMID is frequently accompanied by comorbidities [18], e.g. fibromyalgia syndrome (FMS), osteoarthritis (OA) and other pain-causing diseases (OPCD). This can lead to IMID being unrecognized, as the symptoms (e.g. movement restrictions or joint pain) are put on other diseases which are treated first. However, IMID is the disease among these four which needs most urgent treatment due to possible irreparable consequences. In this case study, we investigate how the recognition of IMID can be supported using statistical classification methods. In particular, we examine whether a diagnosis can be improved by a joint analysis of IMID and its comorbidities, namely FMS, OA and OPCD. In terms of statistical methodology, we ask the question whether the application of MLC constitutes an advantage over SLC in the particular medical context. To this end, we analyze data of patients with suspected or diagnosed rheumatic diseases.

Methods

Classification

Classification is an established supervised learning concept to assign class memberships to samples based on observed variables. In the following, we describe the rationale of classification while highlighting the differences between SLC and MLC. Moreover, we introduce specific classifiers which we use in our case study. More detailed explanations of SLC and MLC are provided in James et al. [23] and Herrera et al. [24], respectively.

Single-label and multi-label classification

We consider tuples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ which consist of covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and target variables $\boldsymbol{y}_i = \left(y_i^1, y_i^2, \ldots, y_i^C\right)$, where $i \in \{1, \ldots, n\}$ indicates the i^{th} sample of a dataset, n is the number of samples, p is the number of covariates, and C is the number of classes. The target variable is a vector $\boldsymbol{y}_i \in \{0,1\}^C$ whose j^{th} component is equal to one if sample i belongs to class j and zero otherwise, i.e. $\forall\ j \in \{1, 2, \ldots, C\}$:

$$y_i^j = \begin{cases} 1, & \text{if sample } i \text{ belongs to class } j \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The vector $\boldsymbol{y}_i$ represents the ground truth and is henceforth called 'true labelset'.

Classifiers aim to suitably map observed covariates $\boldsymbol{x}_i$ to the target; this mapping is based on a model. The classifiers are trained on true labelsets (see Model training and prediction section); the resulting predictions are given as $\hat{\boldsymbol{y}}_i = \left(\hat{y}_i^1, \hat{y}_i^2, \ldots, \hat{y}_i^C\right)$. The entries of this vector are either discrete or continuous, depending on the modeler's choice. Discrete entries are similarly defined as in Eq. (1), i.e. $\forall\ j \in \{1, 2, \ldots, C\}$:

$$\hat{y}_i^j = \begin{cases} 1, & \text{if sample } i \text{ is predicted to belong to class } j \\ 0, & \text{otherwise.} \end{cases}$$

Continuous predictions can e.g. express probabilities or average votes for class memberships, with $\hat{y}_i^j \in [0,1]$ for every j. In the following, we refer to $\hat{\boldsymbol{y}}_i$ as 'predicted labelsets', both for discrete and continuous predictions.

SLC and MLC differ in the fundamental assumption of how many classes a sample can belong to (ground truth). This naturally translates to a difference in the number of labels which a classifier can assign to a sample (prediction). SLC assumes that each sample belongs to exactly one of C classes, i.e. the classes are mutually exclusive. Consequently, $\boldsymbol{y}_i$ is a unit vector so that $\sum_{j=1}^{C} y_i^j = 1$ holds. The latter is also required for $\hat{\boldsymbol{y}}_i$. In discrete prediction, a classifier assigns exactly one of the C labels to each of the samples i (see Model training and prediction section). In case of continuous prediction, the vector $\hat{\boldsymbol{y}}_i$ expresses the probabilities for the classes of sample i. A discrete predicted labelset can still be derived from a continuous one, for example by choosing the label with highest predicted probability. MLC is less restrictive than SLC on ground truth assumptions in the sense that it allows samples to belong to several classes simultaneously, including the option that a sample belongs to none or all of the considered classes. Thus, $\boldsymbol{y}_i$ is not necessarily a unit vector but its components are equal to one for every class j that sample i is part of. Consequently, MLC potentially assigns multiple predicted labels to a sample.

In case of continuous prediction, $\hat{y}_i^j \in [0,1]$ for every j without any further constraint on their sum.

To illustrate application scenarios, we first consider the simplest case of binary classification with $C = 2$ possible outcomes. One may, for example, aim to classify patients as either suffering from a coronary heart attack or not. Here, the considered outcomes 'yes' and 'no' are imperative for SLC; each patient can exclusively belong to one of these classes. In other contexts, however, the two considered classes do not necessarily have to be mutually exclusive. Think about patients who suffer from a coronary heart attack or a pulmonary embolism for instance. Both diseases can occur simultaneously. It depends on the modeler whether the classes are assumed to be mutually exclusive, e.g. due to the low probability of co-occurrence; in that case, one would opt for SLC. Alternatively, the samples may be assumed representative for a population in which simultaneous or no occurrences are considered relevant; this would support the use of MLC.

In both SLC and MLC, the binary case can be extended to more than two classes, i.e. to $C > 2$. Consider a set of patients who suffer from a coronary heart attack, a pulmonary embolism or asthma. There are three classes which they can belong to. As before, one may exclude co-occurrences, resulting in SLC; this is called 'multi-class SLC'. However, it is also possible that none, several or even all of these classes apply to a patient. If we relax the assumption of mutual exclusiveness, a patient can be assigned up to three labels simultaneously, as done in multi-class MLC. In practice, the term 'multi-class' is commonly used in the context of SLC; consequently, the additional terms 'single-label' and 'multi-class' are often only used when particularly relevant.

The different assumptions on simultaneous class membership in SLC and MLC lead to differences in the state spaces of the labelset, the decision spaces and the classification procedures. In Fig. 1, we graphically illustrate the decision and state spaces for $C = 2$ and $C = 3$ classes, both for discrete and continuous prediction and both for SLC and MLC: The state and decision spaces of the C-dimensional labelset appear as a square and a cube, respectively; each edge orientation represents one class, each corner a combination of class assignments from $\{0,1\}^C$. For example, the point (0, 1, 0) of the cube stands for exclusive membership to the second out of three classes; the point (0, 1, 1) represents simultaneous membership to the second and third out of three classes. While MLC allows every combination of class memberships, SLC requires $\sum_{j=1}^{C} y_i^j = 1$. Thus, both the state and decision spaces for SLC are true subsets of those for MLC. For example, the labelsets (0, 0) and (1, 1) can

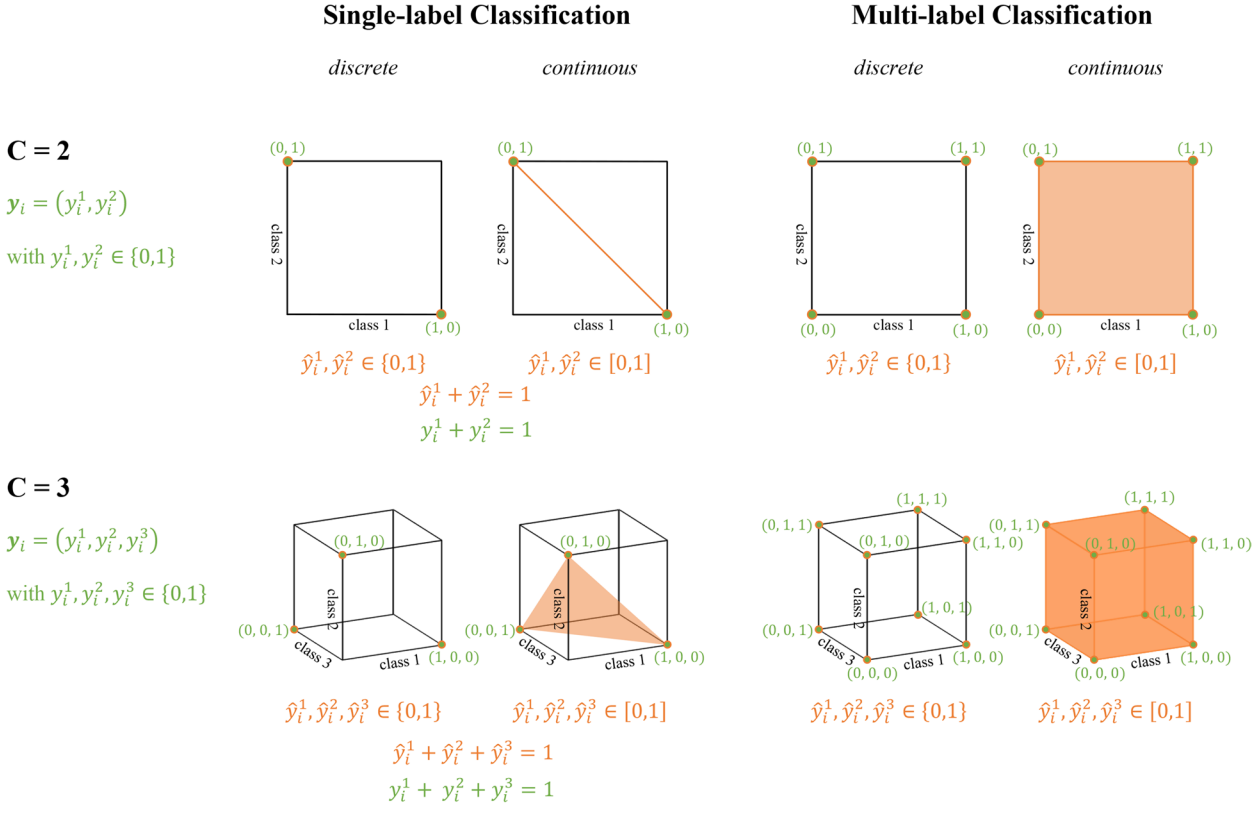

Fig. 1 State spaces of $\boldsymbol{y}_i$ and decision spaces of $\hat{\boldsymbol{y}}_i$ for any sample $i \in \{1, \ldots, n\}$. SLC and MLC are illustrated for $C = 2$ and $C = 3$ classes and for both discrete and continuous prediction. In green, potential true labelsets are depicted, whereas the possible predicted labelsets are shown in orange

only occur in the MLC setting. Assumptions on true class membership naturally extend to the set of possible predictions. Green color stands for the state space of the true labelset $\boldsymbol{y}_i$ (all possible true labelsets), orange color marks the decision space of the predicted labelset $\hat{\boldsymbol{y}}_i$ (all possible predicted labelsets). For discrete predictions, these spaces are identical. For continuous prediction, the decision space expands substantially: For $C = 2$, it stretches to the diagonal (in case of SLC) and the entire surface (MLC); for $C = 3$, it fills a triangular surface (SLC) and the entire cube (MLC).

Model training and prediction

To make meaningful predictions, a classifier needs to learn how to suitably map the covariates $\boldsymbol{x}_i$ to the true labelset $\boldsymbol{y}_i$. To that end, the classifier is trained on a subset of the data (henceforth training set). To assess how accurate the learned mapping classifies new data, it is applied to the remaining subset of the data (henceforth test set) which was not used for model training. By comparing the predicted labelset to the true labelset, the model performance can be evaluated using different measures, e.g. sensitivity, specificity, precision and F1-score. A perfect model, predicting the correct class for each sample, would achieve values of one in these four measures. Thus, the closer the value is to one, the better the model's performance. Subsequently, the learned mapping can be applied to new data which is completely independent of the training and test set (so-called validation set).

A large variety of classifiers is available. In the case study, we use tree-based methods, namely decision trees (DTs) and random forests (RFs), *k*-nearest neighbors (*k*-NN) classifiers, logistic regression models (LRMs) and neural networks (NNs), specifically multi-layer perceptrons (MLPs). The selection is motivated in Methodical procedure section. We roughly outline these methods in the following.

DTs employ step-wise separation of the covariate space into distinct regions [23, 25]. The splitting rules are chosen such that most training samples are correctly classified; the predicted label results as the mean label of all samples falling into such a region. To build an RF, a fixed number of DTs is fitted to the data, each DT making a prediction per sample. The final prediction is then made by majority vote. Majority voting is also used for *k*-NN, considering the *k* nearest points of an observation $\boldsymbol{x}_i$ [25] to assign a class label. The *k* nearest and already classified neighbors are selected using a distance metric, e.g. the Euclidean distance or the Manhattan distance. An LRM is a regression model [23] which connects covariates to a target variable that lies in [0, 1] and represents the probability for a certain class membership. The link between the covariates and target variables is established through the logistic function which maps the real line to the unit interval [0, 1]. Accordingly, the probabilities for the different classes are modeled indirectly via the $\boldsymbol{x}_i$. Regarding NNs, a broadly used type are MLPs [26]. They consist of at least three layers: an input layer, one or more hidden layers and an output layer, each connected to the previous and the following one. NNs predict an outcome by transmitting the input data through all layers by following a feed-forward architecture, i.e. without any loop that allows for a repeated run-through of the layers.

The aforementioned classifiers are directly applicable in the context of SLC. For MLC, the classifiers are combined with one out of three common MLC workflows, namely (i) adaptation-based algorithms, (ii) ensemble-based methods or (iii) transformation-based methods. These are explained in detail in Herrera et al. [24]. Adaptation-based approaches are classifiers, such as a *k*-NN classifier, which are adapted from the single-label to the multi-label setting so that they are able to predict multiple labels simultaneously. Ensemble-based methods combine the results of a set of different single-label classifiers to predict a multi-label vector. Transformation-based methods handle the MLC task by transforming the problem into an SLC task, either into a binary or into a multi-class classification tasks. We apply transformation-based methods in the case study and therefore discuss them in more detail here.

For transformation-based methods, three main approaches can be further distinguished: binary relevance (BR), label powerset (LP) and classifier chain (CC). Fig. 2 graphically visualizes these three approaches. As an example, we consider an MLC problem with $C = 3$ non-exclusive classes. The original dataset is shown at the top of the figure, the three approaches to predict the true labelset $\boldsymbol{y}_i = \left(y_i^1, y_i^2, y_i^3\right)$ for a sample i are illustrated below. For them, covariates are highlighted in gray; exemplary discrete predictions are shown in white.

BR divides the multi-label multi-class task into C independent single-label binary classification tasks and merges the results afterwards (indicated by '+') [27]. In our example, $C = 3$ binary single-label classifiers are used to predict $C = 3$ labels $\hat{y}_i^1, \hat{y}_i^2, \hat{y}_i^3$ per sample i. The set $\left(\hat{y}_i^1, \hat{y}_i^2, \hat{y}_i^3\right)$ forms the predicted labelset. The approach can also be called an ensemble of binary classifiers as it combines BR with the ensemble-based approach [24]. Read et al. [28] state that BR is often seen critically due to the fact that label correlations are unconsidered; however, they highlight the resistance to overfitting and the low computational complexity as convenient properties of BR.

The LP approach transforms the multi-label multi-class task into a single-label multi-class classification problem [24] by increasing the number of classes: Every possible labelset is treated as a new single class, e.g. when

MLC problem: original data

x_{i1}	x_{i2}	x_{i3}	x_{i4}	y_i^1	y_i^2	y_i^3
x_{11}	x_{12}	x_{13}	x_{14}	0	1	1
x_{21}	x_{22}	x_{23}	x_{24}	1	1	1
x_{31}	x_{32}	x_{33}	x_{34}	0	0	1
x_{41}	x_{42}	x_{43}	x_{44}	0	1	1

transformation-based prediction approaches

binary relevance (BR)

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i^1$
x_{11}	x_{12}	x_{13}	x_{14}	0
x_{21}	x_{22}	x_{23}	x_{24}	1
x_{31}	x_{32}	x_{33}	x_{34}	0
x_{41}	x_{42}	x_{43}	x_{44}	1

+

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i^2$
x_{11}	x_{12}	x_{13}	x_{14}	1
x_{21}	x_{22}	x_{23}	x_{24}	1
x_{31}	x_{32}	x_{33}	x_{34}	0
x_{41}	x_{42}	x_{43}	x_{44}	1

+

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i^3$
x_{11}	x_{12}	x_{13}	x_{14}	0
x_{21}	x_{22}	x_{23}	x_{24}	1
x_{31}	x_{32}	x_{33}	x_{34}	1
x_{41}	x_{42}	x_{43}	x_{44}	0

label powerset (LP)

labelset	new label
(0, 1, 1)	1
(1, 1, 1)	2
(0, 0, 1)	3
(0, 1, 1)	1

↓

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i$
x_{11}	x_{12}	x_{13}	x_{14}	1
x_{21}	x_{22}	x_{23}	x_{24}	2
x_{31}	x_{32}	x_{33}	x_{34}	3
x_{41}	x_{42}	x_{43}	x_{44}	2

classifier chain (CC)

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i^1$
x_{11}	x_{12}	x_{13}	x_{14}	0
x_{21}	x_{22}	x_{23}	x_{24}	1
x_{31}	x_{32}	x_{33}	x_{34}	0
x_{41}	x_{42}	x_{43}	x_{44}	1

↓

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i^1$	$\hat{y}_i^2$
x_{11}	x_{12}	x_{13}	x_{14}	0	1
x_{21}	x_{22}	x_{23}	x_{24}	1	1
x_{31}	x_{32}	x_{33}	x_{34}	0	0
x_{41}	x_{42}	x_{43}	x_{44}	1	1

↓

x_{i1}	x_{i2}	x_{i3}	x_{i4}	$\hat{y}_i^1$	$\hat{y}_i^2$	$\hat{y}_i^3$
x_{11}	x_{12}	x_{13}	x_{14}	0	1	0
x_{21}	x_{22}	x_{23}	x_{24}	1	1	1
x_{31}	x_{32}	x_{33}	x_{34}	0	0	1
x_{41}	x_{42}	x_{43}	x_{44}	1	1	0

Fig. 2 Visualization of transformation-based methods BR, LP and CC in the MLC setting to predict the target vector $\boldsymbol{y}_i = \left(y_i^1, y_i^2, y_i^3\right)$. Exemplary original data is given in the blue table at the top. BR fits a binary single-label classifier for each label based on the covariates (gray shaded columns). The resulting predictions (white shaded columns) are combined (visualized by '+') to full labelsets. For LP, the true labelsets are transformed into new classes (orange shaded table); based on this, the multi-label multi-class task turns into a single-label multi-class problem. For CCs, the labels are predicted consecutively. The label order for the prediction in a CC is alterable and may impact the outcome. The example shows discrete prediction; continuous prediction works the same way

originally dealing with $C = 3$ classes, there are $2^C = 8$ possible labelsets and therefore eight classes for the new single-label multi-class problem. This allows the application of well-established SLC techniques. Nonetheless, one big disadvantage of this method is the large number of classes after transformation even for small numbers of classes in the original problem. In addition, the new classes appear with no order, although their content-related closeness varies. But above all, the probability of unobserved labelsets increases with the number of classes, and it is impossible to predict unobserved labelsets using LP [28].

Other than BR, a CC takes into account class correlations as it performs consecutive step-wise binary single-label prediction for the different labels (indicated by the arrows) [28]: CCs start by using the observations $\boldsymbol{x}_i$ to predict the first label $\boldsymbol{y}_i^1$ for every sample, then, it uses the observations $\boldsymbol{x}_i$ and the predictions $\hat{\boldsymbol{y}}_i^1$, which are now considered as truth, as covariates to predict the second label $\boldsymbol{y}_i^2$. Afterwards, the observations $\boldsymbol{x}_i$ and the predictions $\hat{\boldsymbol{y}}_i^1$ and $\hat{\boldsymbol{y}}_i^2$ are used as covariates to predict $\boldsymbol{y}_i^3$. This procedure is continued until $\boldsymbol{y}_i^C$ is reached. For model training, the respective elements of $\boldsymbol{y}_i$ are used in every step of the chain. Special attention should be paid to the order of the label predictions: The order might highly affect the performance of a CC [28] as incorrect predictions in early labels can lead to a bias in the following ones. The use of multiple CCs (referred to as an ensemble of classifier chains, ECC) can help to mitigate the effects of the ordering [28]. An ECC consists of a set of CCs. Each CC is trained on randomly selected (with replacement) samples using a random label order; each CC generates a vector of predictions $\hat{\boldsymbol{y}}_i$ for each *i*. The predictions for the i^{th} sample are subsequently averaged across all CCs and a joint decision is made for the elements of $\hat{\boldsymbol{y}}_i$.

Class imbalance in SLC and MLC

Class imbalance describes the situation of markedly different class frequencies in the training set. Even if this imbalance reflects the characteristics of the underlying population, it leads to practical problems in classification: Common classifiers tend to attach greater importance to the majority class; hence, classifiers use this class as prediction for many samples to increase the number of correctly classified samples and thus achieve satisfying performance measures. Consequently, the minority class is tendentiously ignored [29]. This is problematic in many applications such as fraud detection (e.g. Wei et al. [30]) or disease recognition (e.g. Cohen et al. [31]), where the minority class is of particular interest and its non-consideration might have severe consequences. This makes the issue of class imbalance a relevant and widely discussed topic (e.g. in Guo et al. [29], Sahare and Gupta [32], Gosain and Sardana [33]).

Different methods for handling class imbalance are available. Among these are random over-sampling (ROS, [34]) and random under-sampling (RUS, [34]) as well as synthetic data generation, e.g. via the 'Synthetic Minority Over-sampling Technique' (SMOTE, [35]). The idea behind all these approaches is to artificially equalize class sizes, either by reducing or by increasing them. ROS randomly selects samples (with replacement) of the minority class and replicates them until the number of samples reaches the number of the majority class. As a consequence, ROS entails the risk of overfitting due to duplicates in the data [36]. RUS, in contrast, randomly selects samples of the majority class and eliminates them from the dataset until the classes are of the same size. A drawback of this technique is that the dataset is reduced and potentially informative samples are removed [36]. Synthetic data generation is the process of creating artificial samples for the minority class, which are based on the real data. SMOTE uses the covariates of the *k* nearest minority samples of the original dataset and interpolates between them [35]. This technique avoids the problem of overfitting [36]. Nevertheless, the synthesis of new data is exclusively based on the minority class while the majority class remains unconsidered during this process. Consequently, samples that are ambiguous in terms of their class membership may be generated [37], i.e. synthesized data that should belong to the minority class can simultaneously be close to the majority class.

The described strengths and weaknesses are well known and extensively analyzed in the context of SLC. For MLC, additional difficulties arise, especially when using CCs, where the predictions of labels in the chain are based on the predictions of preceding labels. If these are heavily biased, e.g. due to imbalanced classes, the prediction of the subsequent labels might be biased as well [38]. In addition, the methods described above for handling imbalanced classes are less straightforward to be applied to CCs than in the SLC setting: The true labels might be correlated, and therefore, correcting for an imbalance in one label using, e.g. RUS may cause a new imbalance in a different label. A possible way to handle this is to use a CC with RUS in such a way that RUS is integrated in each step of the CC before predicting the next label [39]. An alternative approach to treat class imbalance in MLC is to apply RUS or ROS at the level of labelsets rather than classes [40]. However, if the difference between frequencies of the most common and the rarest labelset is large, this procedure can lead to the dataset becoming either extremely small (if RUS is applied) or extremely inflated (for use of ROS).

To achieve a balanced compromise, we combine ROS and RUS [41] in our real-data case study: In SLC, classes

with smaller-than-average size are enlarged by ROS, while classes with above-average size are reduced by RUS. The procedure is similar for MLC, with the only difference that ROS and RUS are applied to the *labelsets* rather than *classes*. In the following, we refer to ROS and RUS in the context of SLC as S-ROS and S-RUS, respectively, and in the context of MLC as M-ROS and M-RUS.

Case study

Data

The data analysis was done retrospectively on cross-sectional data collected by the Department of Internal Medicine and Rheumatology at Klinikum Bielefeld, Germany. All patients attended the clinic between 2019 and 2021 as they were suspected to suffer from a rheumatic disease or as they were already diagnosed with it. The original dataset contained observations from 1150 patients of 34 variables; among these were core data (age, sex, admission type, date etc.), information about glucocorticoid intake ('prednisolone') and the CRP value. Further, there were results of different scales to allocate a patient's pain to a certain disease, in particular, results of the Fibromyalgia Rapid Screening Tool (FiRST) [42], the number of tender points as well as results of the Numeric Rating Scale (NRS) [43].

During data preprocessing, we excluded variables related to the course of a disease because we considered these variables subordinate for our research question. In addition, we discarded the body mass index and the weight category which were partially inconsistent with weight and height data. Further, 81 % of both variables were missing. Also for height and weight, the large number of missing values (80 % and 77 %, respectively) was the reason for excluding these variables from our dataset. We further removed information about the examining physician, the admission date and the binary answers to the six single questions of the FiRST questionnaire (while the accumulated score was still included for analysis). The covariates used for analysis are described in more detail in Table 1. The total proportion of missing values in the covariates is 6.4 % and relates to the tender points (28 %), NRS (20 %), the duration of symptoms (10 %), CRP (4 %), the dose of prednisolone (1 %) and the intake of prednisolone (1 %) (see Fig. A1 in the Supplementary Material). We decided to not impute missing values for three reasons: health records are highly individual; particular importance is attached to the empirical covariance structure; and the dataset contains rather few covariables. Therefore, we perform our analysis using complete data of 663 patients.

Each patient underwent investigations for IMID, FMS, OA and OPCD, which distinguishes the case from typical clinical routine data, where exclusion diagnoses are more common. The dataset explicitly documents up to three diagnoses, each assuming to contribute to the symptoms to some extent; the fourth disease can be assumed absent. The main disease, i.e. the disease which contributed most to the symptoms, was recorded as primary diagnosis. The diagnoses represent class assignments made by physicians. We acknowledge that they may deviate from the true underlying disease — particularly as we assume a disease not to be present if a diagnosis is missing. Still, for the purpose of this study, we assume all diagnoses to be correct and therefore constitute the ground truth. Thus, we regard the recorded data as true labelsets, and the classification task will be to derive predicted labelsets.

Based on the noted diagnoses, we derive classes considered by SLC and MLC: For binary SLC, we define the class *IMID* to contain all patients who were diagnosed with IMID as either main disease or as comorbidity. The class contains 257 patients. *noIMID* includes those 406 patients who do not fall into *IMID*. For multi-class SLC,

Table 1 Description of the covariates used for analysis

Variable	Description	Type	Values/ranges	Unit
Age	Age of the patient	numeric	[19; 88]	years
Sex	Sex at birth of the patient	binary	male/female	
Admission type	Admission type	binary	in-/outpatient	
CRP	Measured C-reactive protein	numeric	[0.25; 149.9]	mg/l
Sum FiRST	Sum of positively answered questions of Fibromyalgia Rapid Screening Tool	numeric	[0; 6]	points
Tender points	Number of pain-sensitive pressure points on body	numeric	[0; 18]	points
NRS pain	Numeric rating scale to quantify the pain level	numeric	[0; 10]	points
Prednisolone	Prednisolone intake	binary	yes/no	
Prednisolone dose	Dose of prednisolone intake	numeric	[0.0; 55.0]	mg/day
Duration symptoms	Duration of the symptoms that refer to the main diagnosis	numeric	[1; 600]	months

the class *IMID* remains valid, but we moreover consider the classes *FMS**, *OA** and *OPCD** which we define as the sets of all patients where the respective disease was the main disease but no IMID was diagnosed at all. The class sizes are 168, 61 and 177, respectively. *IMID*, *FMS**, *OA** and *OPCD** are mutually exclusive by definition. The asymmetric consideration of IMID compared to the other diseases is motivated by the main focus of the classification task, namely revealing IMID independently of the rank of the disease. For MLC, we introduce the classes *FMS*, *OA* and *OPCD* as the sets of patients who were diagnosed with the respective disease as either main disease or comorbidity. These classes overlap with *IMID* and with each other. The class sizes are 249, 182 and 298, respectively. Table 2 provides an overview of the considered classes and the approaches in which they are used.

Methodical procedure

We investigate whether the recognition of IMID can be improved by considering comorbidities, even if these are unknown during prediction. In line with the methodological focus of our work, this means that we examine whether MLC recognizes IMID more reliably than SLC. There may be statistical methods apart from the approaches chosen here, especially apart from classification, that might exploit the specific dataset better. However, these would go beyond the scope of our research question of method comparison, so we will limit our analysis to the SLC and MLC methods presented in Classification section. With this design, we make specific methodological choices where they are conceptually justified (e.g. the selection of CC and the focus on IMID), while at the same time conducting a comprehensive assessment of classification performance by systematically varying relevant factors (classifiers, performance measures).

We assume MLC being advantageous over SLC especially if the disease occurrences in our data are associated with each other. Therefore, we first examine the association by performing Pearson's χ^2-test with Yates' continuity correction [44]. To adjust the p-values, we make use of the Bonferroni correction [45]. According to this, we observe an association between IMID and FMS, IMID and OPCD (consisting of the disease groups 'arthralgias of unclear assignment' and 'others'), FMS and OPCD as well as OA and OPCD (see the respective (adjusted) p-values in Table A2 in the Supplementary Material).

We tackle the classification of *IMID* through the following six approaches:

- **Approach 1a: Binary SLC without rebalancing** We consider the two exclusive classes *IMID* and *noIMID* via SLC. In this basic approach, we do not account for class imbalance.
- **Approach 1b: Binary SLC with S-ROS and S-RUS** The class imbalance between *IMID* and *noIMID* in Approach 1a motivates class imbalance correction. Here, we apply a combination of S-ROS and S-RUS. We aim at the average number of samples of both classes in the training set per class as a compromise.
- **Approach 2a: Multi-class SLC without rebalancing** We consider the four exclusive classes *IMID*, *FMS**, *OA** and *OPCD** and tackle the task using multi-class SLC.
- **Approach 2b: Multi-class SLC with S-ROS and S-RUS** The four classes from Approach 2a are highly imbalanced. Consequently, as in Approach 1b, we apply a combination of S-ROS and S-RUS to the training set to account for the class imbalance. Per class, we aim at the average number of samples of all four classes.
- **Approach 3a: Multi-class MLC without rebalancing** As in Approaches 2a and 2b, we consider all four diseases simultaneously. However, we now employ the non-exclusive classes *IMID*, *FMS*, *OA* and *OPCD*. In other words, we allow patients to have multiple labels simultaneously and consequently perform MLC. Figure 3 and Table A1 in the Supplementary Material display the number

Table 2 Description of the different classes considered for SLC and MLC. The resulting labelsets for the MLC approach and their frequencies are visualized in Fig. 3 and provided in tabular form in Table A1 in the Supplementary Material

Class	Description	Approach	Class size
IMID	diagnosed with IMID as main disease or comorbidity	binary SLC, multi-class SLC, MLC	257
noIMID	no IMID diagnosed	binary SLC	406
*FMS**	diagnosed with FMS as main disease, no IMID diagnosed at all	multi-class SLC	168
*OA**	diagnosed with OA as main disease, no IMID diagnosed at all	multi-class SLC	61
*OPCD**	diagnosed with OPCD as main disease, no IMID diagnosed at all	multi-class SLC	177
FMS	diagnosed with FMS as main disease or comorbidity	MLC	249
OA	diagnosed with OA as main disease or comorbidity	MLC	182
OPCD	diagnosed with OPCD as main disease or comorbidity	MLC	298

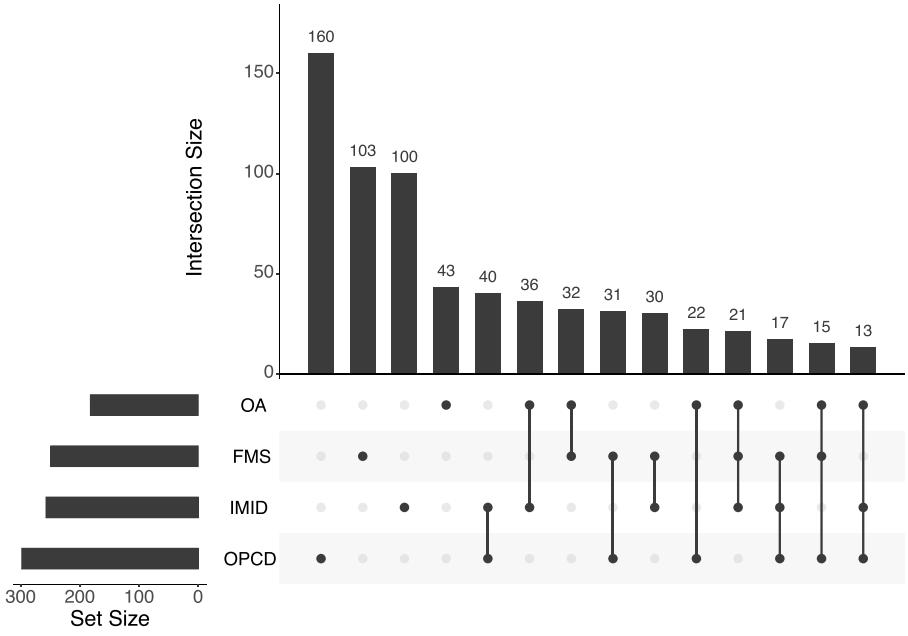

Fig. 3 Number of (joint) occurrences of the four diseases in the case study dataset. Each patient suffers from at least one disease; no patient suffers from all four diseases. The true labelsets (0, 0, 0, 0) and (1, 1, 1, 1) are not present in the data. A tabular representation of all numbers is additionally provided in Table A1 in the Supplementary Material

of occurrences of true labelsets. To handle the classification task, we use the transformation-based CCs (see Model training and prediction section). BR is discarded because of its inability to consider label correlations. Further, we abstain from using LP as it is unsuited to predict unobserved labelsets [28]. In contrast, CCs allow us to investigate whether the prediction of FMS, OA and OPCD improves the recognition of IMID from the present data. Although only 14 out of 16 labelsets are represented in the data (ground truth), all labelsets can be predicted due to the step-wise binary prediction of the four labels using CCs. Within the CC approach, we consider different label orderings. However, as our research question implies that IMID should be predicted as last element of the target vector, we estimate the MLC models assessing only six different label orderings instead of 24. In Results section, we report and discuss the results of only that ordering which leads to the highest average F1-score for the majority of all applied classifiers; that is the ordering where y_i^1 corresponds to FMS, y_i^2 to OPCD, y_i^3 to OA and y_i^4 to IMID. For the sake of complete reporting, we visualize the results of the other label orderings in Fig. A3 in the Supplementary Material; they differ only slightly.

- **Approach 3b: Multi-class MLC with M-ROS and M-RUS** We perform MLC with the same classes as in Approach 3a, this time with a combined procedure of M-ROS and M-RUS to account for class imbalance. In contrast to the proceeding in the context of SLC (S-ROS and S-RUS, Approach 2b), we apply the rebalancing method to the true labelsets rather than to the classes (which were defined differently, and thus the procedures are not transferable to each other; see Class imbalance in SLC and MLC section for further details). The smallest class among the 14 present labelsets is the group of patients who simultaneously suffer from OPCD, OA and IMID and consists of 13 samples. The largest class contains 160 samples; these patients exclusively suffer from OPCD. Using a combination of M-ROS and M-RUS, we aim at the average number of samples of all true labelsets in the training set.

For each approach, we apply DT, RF, LRM, *k*-NN and MLP as classifiers in order to take various requirements into account. DT, *k*-NN and LRM are easily interpretable, which is an important criterion for the use of statistical models in the medical field. The rather complex structure of medical data due to the individuality of patients, however, might require more complex models. Hence, we use RF and MLP as classifiers and accept their reduced interpretability. To account for the rather small amount of data and avoid overfitting of the models, we apply nested cross-validation (CV). Hence, we validate the models' performance on different training and test splits. The outer loop is a ten-fold CV, the inner loop is a five-fold CV to optimize the hyperparameters of the classification

models using a grid search. This design implies that numbers of instances that are created or removed through RUS or ROS differ across the folds of the nested CV, whilst previously indicated class sizes refer to the entire dataset before splitting it into training and test data. As the F1-score is the harmonic mean of the sensitivity and the precision, we use macro F1-score to evaluate each of the cross-validated models, and we refit the optimized model using the same measure. Due to the need for a rapid initiation of a treatment in case of the IMID, the sensitivity and the precision are of higher priority than the specificity. Reported results are class-specific for IMID. For the DTs, we tune the criterion to measure the quality of a split, the maximum depth of the tree and the minimum number of samples in a node to perform a further split. In addition to these hyperparameters, we optimize the number of trees for the RFs. For the LRMs, we tune the penalty norm, the tolerance for the stopping criteria as well as the used algorithm for solving the optimization problem. When estimating the MLPs, we optimize the number of hidden layers each consisting of 100 neurons, the activation function of the hidden layer, the solver, the strength of the $L2$ regularization term and the learning rate. In case of the k-NN classifier, we tune the number of considered neighbors k. Recall Fig. 3 for a visualization of the number of occurrences of the four diseases in the dataset and their intersections. A tabular form of these numbers is additionally provided in Table A1 in the Supplementary Material.

Results

In the following, we outline the results of the application of all models described in Methodical procedure section to the data from Data section. The IMID-specific performance per approach and per classifier across all ten CV folds is shown in Fig. 4; a tabular representation of the averaged results and the respective 95 % confidence intervals are provided in Table A3 in the Supplementary Material. Receiver operating characteristic curves are provided in Fig. A2 in the Supplementary Material. The focus on prediction performance of IMID rather than any other disease or disease combination arises from the case study's objectives.

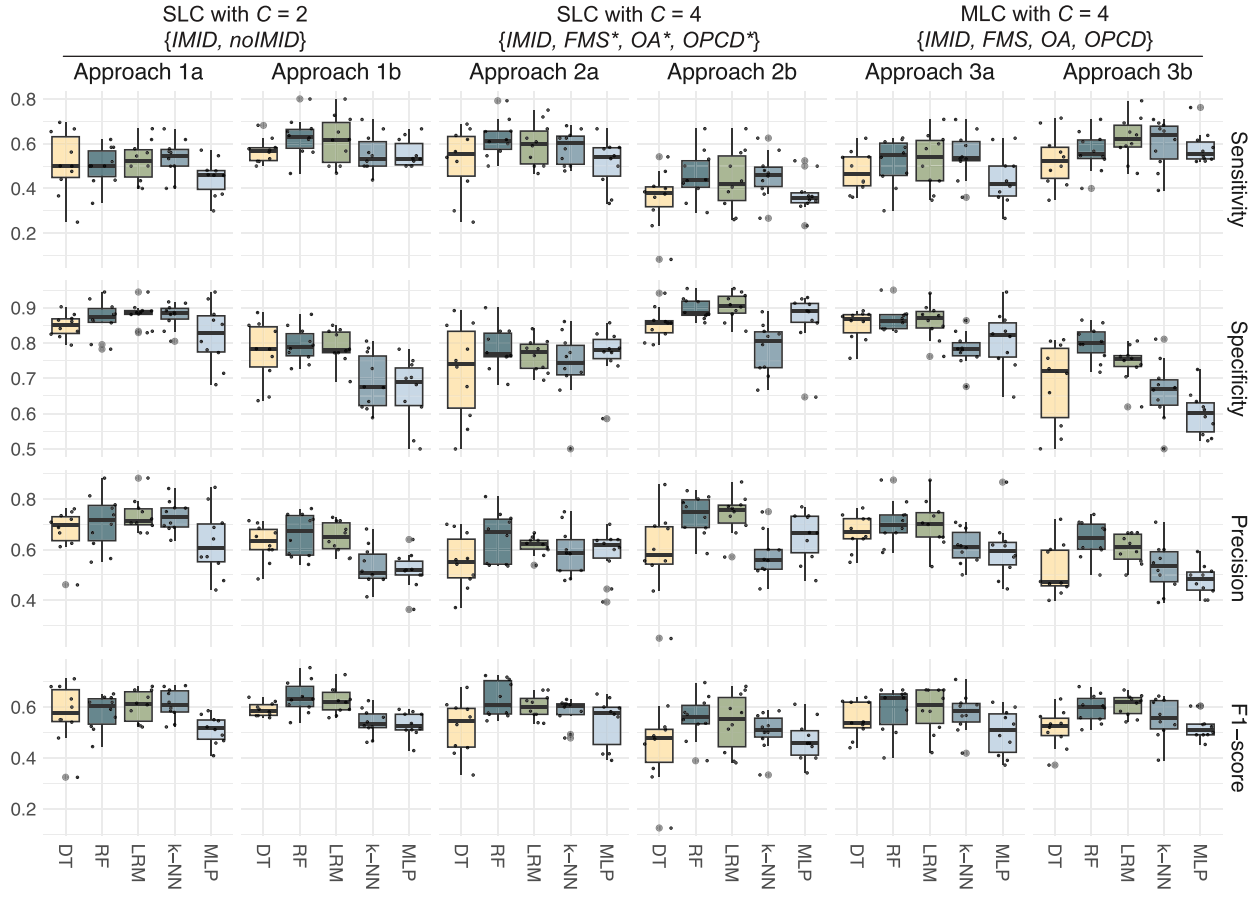

Fig. 4 IMID-specific performance of classification models for the six described approaches. A tabular representation of the averaged results and the respective 95 % confidence intervals are provided in Table A3 in the Supplementary Material. Approaches with suffix *a* are without rebalancing of classes or labelsets, while approaches with suffix *b* are with rebalancing

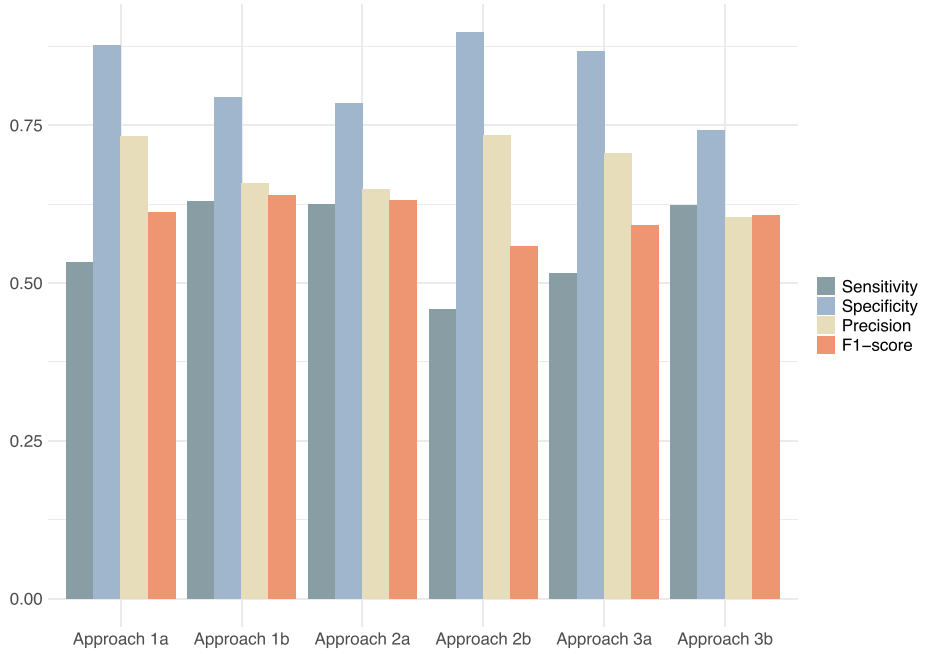

Fig. 5 Sensitivity, specificity, precision, and F1-score for the best-performing classifier for each approach, based on F1-score if best scores are spread out across different classifiers. For Approaches 1b, 2a, 2b and 3a, RF is best, whereas MLP is best for Approach 1a and LRM is best for Approach 3b

Comparison of classifiers

Firstly, we compare the classifiers DT, RF, LRM, *k*-NN and MLP with each other by looking at the approaches individually, i.e. column-by-column in Fig. 4.

For **Approach 1a**, the four best measures are spread out across two classifiers, namely LRM and MLP. MLP performs best regarding the sensitivity (53.3 %) and the F1-score (61.2 %), whereas LRM achieves the highest precision (73.5 %) and the highest specificity (88.3 %). The two latter values are good results; however, the remaining two measures are too low for a reliable classification. The results are quite similar for all classifiers; *k*-NN achieves the lowest value in all four measures.

In **Approach 1b**, RF performs best regarding all measures: It results in a sensitivity of 62.9 %, a specificity of 79.5 %, a precision of 65.8 % and an F1-score of 63.9 %. All measures are moderately high, but still too low for a valid use in the medical field.

The highest results of **Approach 2a** are almost identical to those of Approach 1b: RF achieves a sensitivity of 62.5 %, a specificity of 78.5 %, a precision of 64.8 % and an F1-score of 63.2 %. MLP and LRM achieve comparable results, DT and *k*-NN perform worse.

Approach 2b leads to a low sensitivity (at most 45.9 %, achieved by RF) as well as to a low F1-score (at most 55.8 %, achieved by RF), but a high specificity (90.4 %, achieved by LRM) and a moderately high precision (74.0 %, achieved by LRM); this pattern is recognizable for RF and LRM, whereas DT, MLP and *k*-NN result in low measures except from the specificity.

Approach 3a results in a low sensitivity (55.0 %, achieved by MLP) and a low F1-score (59.1 %, achieved by RF). The precision is moderately high at 70.6 %; the specificity, in contrast, is high at 86.7 % (both achieved by RF).

As in Approaches 1b and 2a, the models of **Approach 3b** show a moderate performance; viewed in isolation, almost all classifiers perform rather weak regarding the sensitivity. LRM results in the highest sensitivity of 62.3 %; the highest F1-score of 60.7 % achieved by LRM is rather low. Even the specificity of 80.0 % achieved by RF is, compared to other approaches, a weak result. Across all classifiers, the precision is rather low; it ranges from 47.9 % achieved by *k*-NN to 64.0 % achieved by RF.

To draw an interim conclusion: The best performances of the six approaches is spread out across RF, LRM and MLP, with a noticeable accumulation at RF. In contrast, DT and *k*-NN show comparably low performance. Measured by the F1-score, the RF seems to be the best suited classifier for our analysis in the majority of approaches; in more detail, RF is best for Approaches 1b, 2a, 2b and 3a, whereas MLP is best for Approach 1a, and LRM is best for Approach 3b. Sensitivity, specificity, precision, and F1-score of the best-performing classifier for each approach is visualized in Fig. 5.

Comparison of approaches

When weighing up the six Approaches 1a to 3b against each other, we take particular account of row-wise

comparisons in Fig. 4. In the following, we go into more detail about the differences between the six approaches.

Approaches with vs. without rebalancing: We first compare the approaches in pairs, namely those accounting for the same number of classes and allowing for the same number of labels, but one with and one without rebalancing (that is: a vs. b). After this, we highlight the differences between the three approaches without rebalancing (a), followed by a comparison of those accounting for class imbalance (b). Due to the importance of detecting patients suffering from IMID as well as being precise in the recognition, we pay special attention to the F1-score for the purpose of selecting the best performing approach.

Compared to the results of **Approach 1a**, the F1-score of all classifiers, except from MLP, in **Approach 1b** increased by 1.6 to 5.9 percentage points (pp); the F1-score of MLP is decreased by 7.3 pp. For most classifiers, the sensitivity increases to a considerable extent (by max. 12.8 pp), but at the same time the specificity and the precision decrease (by max. 18.7 pp and 19.7 pp, respectively). The difference between these two approaches lies in the handling of class imbalance. While the classes in Approach 1a are highly imbalanced (the entire dataset contains 257 patients with IMID and 406 patients without), both classes are equally represented in Approach 1b. Considering this class imbalance, the low sensitivity and high specificity of the models in Approach 1a are reasonable, as the positive samples (i.e. the patients suffering from IMID) are underrepresented. Consequently, the changes in results meet our expectations of an increased sensitivity and a decreased specificity. Albeit an increase in sensitivity is desirable (the models get better at predicting patients who require rapid treatment as they suffer from IMID), a decrease in specificity might lead to an overstrain of clinics resulting in long waiting times which increase the time to first treatment. Nevertheless, with regard to the F1-score, Approach 1b mostly outperforms Approach 1a (see Fig. 5).

Approaches 2a and 2b consider $C = 4$ classes each. Comparing these two approaches, Approach 2b results in a considerably worse sensitivity (decreased by up to 16.6 pp) but in a strongly increased specificity (by max. 13.9 pp) for all classifiers. This is similar for the precision and the F1-score: The F1-score decreases (my max. 8.4 pp), while the precision increases (by max. 12.4 pp). Only for MLP, the precision decreases by 1.4 pp. In the entire dataset, the difference of 196 samples between the largest (IMID) and the smallest class (OA) is large. Accordingly, the decrease in sensitivity and the increase in specificity were to be expected, as IMID, the class of main interest, is downsampled for the purpose of class rebalancing which reduces the amount of potentially useful information. Nevertheless, without handling class imbalance, our model might be biased and would consequently result in too many false positively predicted samples. The rebalancing, however, leads to a strong decrease in sensitivity which results in a non-identification of patients who need rapid treatment. When selecting an approach based on the F1-score, Approach 2a outperforms Approach 2b (see Fig. 5).

In **Approaches 3a and 3b**, we utilize the fact that FMS, OA and/or OPCD frequently accompany IMID; we there analyze the data using MLC. Rebalancing in the context of these models increases the sensitivity by 4.7 to 13.6 pp, but lowers the specificity by 6.7 to 20.8 pp. At the same time, the precision decreases by up to 14.8 pp. The F1-score yields an ambiguous picture, ranging from a decreases by at max. 3.7 pp to an increase by up to 2.1 pp. Similar to the difference between Approaches 1a and 1b, the increased sensitivity allows for a better recognition of those patients suffering from IMID; however, the specificity is too low to certainly identify patients without IMID. Nevertheless, Approach 3b outperforms Approach 3a based on the F1-score (see Fig. 5).

As an interim conclusion, we summarize that the difference in model performance between unbalanced and balanced approaches depends on (i) the number of considered classes and (ii) the consideration as either SLC or MLC problem. For $C = 2$ (Approaches 1a and 1b), SLC performs better at predicting IMID after rebalancing, although the sensitivity is insufficient for a disease that requires rapid treatment. For $C = 4$, however, rebalancing worsens the recognition of IMID for SLC (Approaches 2a and 2b) but improves it for MLC (Approaches 3a and 3b). This makes sense as IMID class sizes vary across approaches; consequently, it is sometimes upsampled, sometimes downsampled, resulting in varying sensitivity.

In the following, we compare the three approaches (i.e. (1a/b) binary SLC, (2a/b) multi-class SLC, and (3a/b) multi-class MLC) with each other, first without rebalancing, second with rebalancing.

Approaches without rebalancing: Approaches 1a, 2a and 3a differ in the aforementioned two main points: (i) the number of considered classes and (ii) the consideration as either SLC or MLC problem. Regarding (i), Approach 1a considers $C = 2$ classes, whereas Approaches 2a and 3a both distinguish $C = 4$ classes. Regarding (ii), Approaches 1a and 2a use SLC, whereas Approach 3a uses MLC. MLC allows a more generous assignment of IMID labels, since the occurrence of other diseases does not exclude this. In addition, considering comorbidities for the recognition of IMID increases the set of possibly informative covariates compared to SLC. Still, the sensitivity of all models belonging to Approach 3a are worse than the ones of Approach 2a; even compared to Approach 1a, the sensitivity achieved

by DT and *k*-NN is lower. In case of RF, for instance, the sensitivity is 10.9 pp lower than in Approach 2a and just 1.5 pp higher than in Approach 1a. A possible explanation for the decreased sensitivity in Approach 3a is that the prediction of IMID is based on the prediction of the other three labels due to the construction of CCs. However, the precedent predictions might be imprecise due to highly imbalanced classes. As a result, the prediction of IMID might be imprecise as well.

Approaches with rebalancing: As motivated above, handling class imbalance is necessary to aim at precise models, which is why we applied rebalancing methods to our data. Among the models with rebalancing (Approaches 1b, 2b and 3b), the poor sensitivity of Approach 2b is immediately apparent. Compared to the Approaches 1b and 3b, it is at max. 21.1 and 20.9 pp, respectively, lower. The specificity, however, is highest in Approach 2b for every classifier. This also applies to the precision, except for DT. Regarding the F1-score, Approach 1b performs best in case of DT, RF, LRM and *k*-NN; in case of MLP, Approach 3b is best. Based on these results and assuming the F1-score being the most important measure, Approach 1b seems to perform best on our data (see Fig. 5). As for the unbalanced approaches, the MLC methods do not generally outperform SLC although IMID is frequently accompanied by FMS, OA and/or OPCD, and although Pearson's χ^2-test indicated an association between the diseases (see Table A2 in the Supplementary Material), except for IMID and OA as well as for FMS and OA. As an additional analysis, we applied MLC without consideration of the class OA, but this lead to no noticeable difference to the reported models (see Fig. A4 in the Supplementary Material).

Discussion

Our empirical results present an ambiguous picture: We did not find clear evidence that analyzing IMID in isolation (Approach 1a/b) leads to improved diagnostic performance when compared with a more differentiated consideration of diseases (Approach 2a/b) or the inclusion of comorbidities (Approach 3a/b). While the overall best IMID-specific sensitivity and F1-score is achieved by Approach 1b (62.9 % and 63.9 %, respectively), the overall best specificity and precision is reached by Approach 2b (90.4 % and 74.0 %, respectively). However, performance varies across approaches, application of rebalancing and classifiers; the only method that never outperforms the others (for no classifier and none of the four measures) is Approach 3a, i.e. MLC without rebalancing.

With regard to our study, weak performance or the fact that taking comorbidities into account does not lead to a clear improvement may be due to the following factors:

- **Correlation strength:** The association between the diseases may be too weak for an added value of considering comorbidities. Even if disease occurrence is correlated, this may be insufficient to provide relevant information for the recognition of IMID.
- **Missing information:** The used data may miss relevant covariates for reliable recognition of IMID. It contains 13 variables, but most of them are rather related to FMS. This applies, in particular, to the FiRST questionnaire and the tender points. The CRP value is a valid indicator for any inflammation in the body; consequently, the variable is not necessarily sufficient to clearly separate the diseases from one another. For OA in particular, no typical variables (e.g. X-ray findings or other variables which query frequently occurring symptoms) are included in the dataset. This makes the recognition of comorbidities and thus the recognition of IMID more difficult.
- **Class imbalance:** The disparity in class sizes is partly substantial. This imbalance may bias the results: Our class imbalance correction often improves the classifiers' performance, but the dataset is strongly manipulated by excluding and duplicating samples.
- **Selection of patients:** Some of the patients in the dataset were non-recently diagnosed and already received treatment before data collection. This affects, for example, the CRP value so that it no longer corresponds to the value without treatment; this makes clear pattern recognition more difficult. Moreover, we reduced the dataset to complete cases for good reasons. However, values were possibly not missing at random, i.e. their absence might have contained implicit information. For example, health records of patients with a more complicated course or an ambiguous clinical picture might have been documented in more detail. This may have affected the predictive performance of the models.

Despite the inconclusive picture and the limitations of the dataset, the data application illustrates the influence that different numbers of classes and the interpretation as an SLC or MLC problem can have. It also shows how the correction for class imbalance might affect the results. In doing so, it demonstrates the challenges of disease classification using a real-world dataset.

Conclusion

In this work, we provided a methodological examination of SLC and MLC approaches, with a particular focus on disease recognition based on health records. By explaining and comparing both frameworks conceptually and empirically, we illustrated how underlying assumptions, learning structures, and strategies for handling class imbalance shape predictive performance. Using

patient-derived data from a clinical setting, we demonstrated the practical implications of these methodological choices, particularly regarding the simultaneous consideration of comorbidities. In doing so, our study connects conceptual clarification with empirical validation, thereby contributing to a more systematic understanding of when and how SLC or MLC may be advantageous in medical classification tasks.

In a case study, we investigated whether the simultaneous consideration of comorbidities results in a more reliable recognition of pain-causing diseases. The aim of the analyses was to compare SLC and MLC, together with other choices in connection with the training and prediction procedure, rather than to identify the best possible model for recognizing the disease of interest. In detail, we performed a case study on 663 patients with suspected or diagnosed rheumatic diseases, namely IMID, FMS, OA and OPCD. The study investigated the effect of considering different numbers of classes, and of interpreting the problem as SLC or MLC. In the context of MLC, we chose CC, a transformation-based method, in order to account for label correlations. For both SLC and MLC, class imbalance was handled using a combination of ROS and RUS. To examine the impact of these procedural decisions, we applied six different approaches to the same data. Each of them used DTs, RFs, LRMs, *k*-NN and MLP as classifiers.

The analyses yielded an ambiguous picture. Each method choice entailed different strengths and weaknesses: between SLC or MLC, between the number of classes, the correction for class imbalance, the classification method, and the evaluation measure. Modeling decisions are (partly inevitably) entangled, so that the effect of SLC versus MLC cannot be completely isolated. This underlines the importance of addressing each individual problem in an equally individual manner, especially when the implications can be significant, such as in patient referrals: How strongly are interdependencies between diseases assumed? How complex (or, e.g. non-linear) does the relationship between disease and covariates appear, and to what type of statistical classification method does this lead? Which aspect of performance do we consider most important, e.g. sensitivity or specificity?

A look at the situation underlying the examined use case helps to, for example, contextualize the relevance of different performance measures: Current practice shows that (too) many suspected cases are referred to rheumatology departments, where the diagnosis is often not confirmed by specialists. This unnecessarily prolongs waiting times for patients who are actually affected by IMID. At the same time, it remains crucial that every patient has the opportunity to be examined in a specialized clinic.

The results of our study lead to the point that computational prediction tools should always serve as decision *support* for treating physicians: complementing, but never replacing, their clinical judgment. Against this background, the question arises of how classification results should be presented. In our use case, we applied hard (discrete) classification thresholds, predicting patients as either diseased or non-diseased, as this approach suited the research question of comparing the performance between methods. In the context of decision support, however, it is immediately possible to generate continuous predictions, such as probabilities. Each of the approaches discussed here is capable of providing such information. Moreover, the moderate predictive performance of our models prompts to the question as to how good the performance needs to be for a model to support physicians in their decision-making process. No general statement can be made here as the minimum performance required depends heavily on the context: While a high sensitivity is most crucial for life-threatening diseases, one may pay larger attention to the sensitivity for less severe diseases to avoid unnecessary referrals. Further, the required level of confidence of the system must be considered in conjunction with the expertise of the physician making the decision. The comprehensibility of models is key, even if a model supports the physician's preliminary decision.

Independent of the specific use case, the analyzed data, and the observed performance results, it is important to consider the conceptual advantages of MLC compared to SLC. While SLC requires a decision for a single class, even when the available covariates may indicate multiple possible diseases, MLC allows for the simultaneous assignment of several conditions to one patient. This reflects a more realistic representation of clinical practice, where multiple diseases can coexist and should therefore be jointly considered in diagnostic support. At the same time, it is important to emphasize that neither approach conceptually dominates the other, neither in theory nor in our empirical study: In practice, models with a larger number of classes do not necessarily yield better results than those with fewer classes, and MLC is not inherently superior to SLC. In practical implementation, computational costs and the effort required for correcting class imbalance —both of which tend to increase with a higher number of classes and when using MLC — also need to be carefully weighed. Overall, when selecting either the SLC or the MLC approach, we recommend to critically examine the circumstances of the particular application: That is, assumptions on the ground truth (causal relations between or common co-occurrences of diseases, expected effect sizes, number of considered diseases), characteristics of available data (missing diagnoses, availability of covariates) and the purpose of the analysis (inference vs. prediction). For example, if comorbidities are strongly correlated with a disease of main interest,

MLC might be beneficial for its recognition, especially if the diseases present with antagonistic symptoms. However, if the dataset misses relevant disease-specific covariates, SLC may be computationally advantageous. Regarding small sample sizes or class imbalance, none of the approaches is superior to the other. The interpretability of the models depends in particular on the choice of classifier, both for SLC and MLC. An informed and well-founded decision can be made on the basis of synthetic data which is tailored to explicit assumptions about the ground truth. Such data can be generated either specifically for the use case at hand, as proposed by Marchi et al. [46], or in a more general framework, as for example described by Kotelnikov et al. [47].

In summary, our study highlights the potential of statistical decision support to improve clinical decision-making. The underlying methods extend well beyond the health context. Their usefulness increases with the availability of high-quality data and the careful selection of appropriate statistical approaches: The better the alignment between the use case, the data, and the methodological assumptions, the more reliable and practically relevant the results become. Future research should therefore focus on refining data quality, exploring advanced modeling techniques, and developing transparent, interpretable tools that can effectively assist physicians — and professionals in other domains — in making informed decisions.

Abbreviations

BR	Binary relevance
CC	Classifier chain
CRP	C-reactive protein
CV	Cross-validation
DT	Decision tree
ECC	Ensemble of classifier chains
FiRST	Fibromyalgia Rapid Screening Tool
FMS	Fibromyalgia syndrome
IMID	Immune-mediated inflammatory diseases
LP	Label powerset
LRM	Logistic regression model
k-NN	*k*-nearest neighbors
MLC	Multi-label classification
MLP	Multi-layer perceptron
NN	Neural network
NRS	Numeric Rating Scale
OA	Osteoarthritis
OPCD	Other pain-causing diseases
RF	Random forest
ROS	Random over-sampling
RUS	Random under-sampling
SLC	Single-label classification
SMOTE	Synthetic Minority Over-sampling Technique

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-026-02934-w.

Additional file 1. Additional information on the dataset used for analyses and on the results of the case study is provided in the Supplementary Material.

Acknowledgements

We thank Tamara Schamberger, Philipp Hege and Valentina Lakusta for helpful comments and discussions.

Code availability

The code used for model fitting is available at https://github.com/fuchslab/SLC_MLC_Disease_Recognition.

Authors' contributions

S.S. conceptualized the study, analyzed and visualized the data, and was a major contributor in writing the manuscript. H.M. conceptualized the study. M.-H.R. was involved in data curation. M.R. was involved in data curation and funding acquisition. C.F. conceptualized and supervised the study, and was involved in funding acquisition. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Medical Research Start-up Fund of the Medical School OWL, Bielefeld University.

Data availability

Due to data privacy protection, we are not allowed to share the original data. Data preprocessing was performed using R version 4.4.2 [48]. The SLC models as well as the MLC models were fitted using Python version 3.12.13 [49].

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Our study adhered to the Declaration of Helsinki. Ethical approval was given by the Ethics Committee of the Westphalia-Lippe Medical Association and the University of Münster (Ethik-Kommission der Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität Münster); case number: 2021-426-f-S. Informed consent was obtained from all participants or their legal guardian.

Consent for publication

Not applicable.

Received: 6 March 2026 / Accepted: 26 June 2026
Published online: 10 July 2026

References

1. Dangare CS, Apte SS. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. Int J Comput Appl. 2012;47(10):44–8. https://doi.org/10.5120/7228-0076.
2. Vijayarani S, Dhayanand S. Data Mining Classification Algorithms for Kidney Disease Prediction. Int J Cybern Informat. 2015;4(4):13–25. https://doi.org/10.5121/ijci.2015.4402.
3. Gárate-Escamila AK, Hassani HE, Andrès A, E. Classification Models for Heart Disease Prediction Using Feature Selection and PCA. Inform Med Unlocked. 2020;19:100330. https://doi.org/10.1016/j.imu.2020.100330.
4. Kwon OW, Lee JH. Text Categorization Based on k-nearest Neighbor Approach for Web Site Classification. Inf Process Manag. 2003;39(1):25–44. https://doi.org/10.1016/s0306-4573(02)00022-5.
5. Crammer K, Dredze M, Ganchev K, Talukdar PP, Carroll S. Automatic Code Assignment to Medical Text. In: Biological, translational, and clinical language processing. 2007. p. 129–36. https://dl.acm.org/doi/10.5555/1572392.1572416.
6. Liu J, Chang WC, Wu Y, Yang Y. Deep Learning for Extreme Multi-label Text Classification. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2017. p. 115–24. https://doi.org/10.1145/3077136.3080834.
7. Dineva A, Mosavi A, Gyimesi M, Vajda I, Nabipour N, Rabczuk T. Fault Diagnosis of Rotating Electrical Machines Using Multi-Label Classification. Appl Sci. 2019;9(23, 5086). https://doi.org/10.3390/app9235086.

8. Turnbull D, Barrington L, Torres D, Lanckriet G. Semantic Annotation and Retrieval of Music and Sound Effects. IEEE Trans Audio Speech Lang Process. 2008;16(2):467–76. https://doi.org/10.1109/tasl.2007.913750.
9. Trohidis K, Tsoumakas G, Kalliris G, Vlahavas I. Multi-label Classification of Music by Emotion. EURASIP J Audio Speech Music Process. 2011;2011:4. https://doi.org/10.1186/1687-4722-2011-426793.
10. Abdel Maksoud EA, Barakat S, Elmogy M. 9. In: Medical Images Analysis Based on Multilabel Classification. Elsevier; 2019. p. 209–245. https://doi.org/10.1016/b978-0-12-816086-2.00009-6.
11. Farooq A, Anwar S, Awais M, Rehman S. A deep CNN based Multi-class Classification of Alzheimer's Disease using MRI. In: 2017 IEEE International Conference on Imaging Systems and Techniques (IST). IEEE; 2017. https://doi.org/10.1109/ist.2017.8261460.
12. Wosiak A, Glinka K, Zakrzewska D. Multi-label Classification Methods for Improving Comorbidities Identification. Comput Biol Med. 2018;100:279–88. https://doi.org/10.1016/j.compbiomed.2017.07.006.
13. Zufferey D, Hofer T, Hennebert J, Schumacher M, Ingold R, Bromuri S. Performance Comparison of Multi-label Learning Algorithms on Clinical Data for Chronic Diseases. Comput Biol Med. 2015;65:34–43. https://doi.org/10.1016/j.compbiomed.2015.07.017.
14. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II: A Public-access Intensive Care Unit Database*. Crit Care Med. 2011;39(5):952–60. https://doi.org/10.1097/ccm.0b013e31820a92c6.
15. Sajid NA, Rahman A, Ahmad M, Musleh D, Basheer Ahmed MI, Alassaf R, et al. Single vs. Multi-Label: The Issues, Challenges and Insights of Contemporary Classification Schemes. Appl Sci. 2023;13(11, 6804). https://doi.org/10.3390/app13116804.
16. Michielan L, Terfloth L, Gasteiger J, Moro S. Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates. J Chem Inf Model. 2009;49(11):2588–605. https://doi.org/10.1021/ci900299a.
17. Wu S, Chen Y, Li Z, Li J, Zhao F, Su X. Towards Multi-label Classification: Next Step of Machine Learning for Microbiome Research. Comput Struct Biotechnol J. 2021;19:2742–9. https://doi.org/10.1016/j.csbj.2021.04.054.
18. McInnes IB, Gravallese EM. Immune-mediated Inflammatory Disease Therapeutics: Past. Present Futur Nat Rev Immunol. 2021;21:680–6. https://doi.org/10.1038/s41577-021-00603-1.
19. Anaya JM, Gómez L, Castiblanco J. Is there a Common Genetic Basis for Autoimmune Diseases? J Immunol Res. 2006;13(2–4):185–95. https://doi.org/10.1080/17402520600876762.
20. Rahman P, Inman RD, El-Gabalawy H, Krause DO. Pathophysiology and Pathogenesis of Immune-mediated Inflammatory Diseases: Commonalities and Differences. J Rheumatol Suppl. 2010;85:11–26. https://doi.org/10.3899/jrheum.091462.
21. Russell AS, Gulliver WP, Irvine EJ, Albani S, Dutz JP. Quality of Life in Patients with Immune-mediated Inflammatory Diseases. J Rheumatol Suppl. 2011;88:7–19. https://doi.org/10.3899/jrheum.110899.
22. Pope JE, Choy EH. C-reactive Protein and Implications in Rheumatoid Arthritis and Associated Comorbidities. Semin Arthritis Rheum. 2021;51(1):219–29. https://doi.org/10.1016/j.semarthrit.2020.11.005.
23. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. 2nd ed. Springer US; 2021. https://doi.org/10.1007/978-1-0716-1418-1.
24. Herrera F, Charte F, Rivera AJ, del Jesus MJ. Multilabel Classification: Problem Analysis, Metrics and Techniques. Springer eBook Collection. Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-41111-8.
25. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7.
26. Popescu MC, Balas VE, Perescu-Popescu L, Mastorakis N. Multilayer Perceptron and Neural Networks. WSEAS Trans Circ Syst. 2009;8(7):579–88.
27. Zhang ML, Li YK, Liu XY, Geng X. Binary Relevance for Multi-label Learning: An Overview. Front Comp Sci. 2018;12(2):191–202. https://doi.org/10.1007/s11704-017-7031-7.
28. Read J, Pfahringer B, Holmes G, Frank E. Classifier Chains for Multi-label Classification. Mach Learn. 2011;85:333–59. https://doi.org/10.1007/s10994-011-5256-5.
29. Guo X, Yin Y, Dong C, Yang G, Zhou G. On the Class Imbalance Problem. In: Fourth International Conference on Natural Computation. IEEE; 2008. p. 192–201. https://doi.org/10.1109/icnc.2008.871.
30. Wei W, Li J, Cao L, Ou Y, Chen J. Effective Detection of Sophisticated Online Banking Fraud on Extremely Imbalanced Data. World Wide Web. 2013;16:449–75. https://doi.org/10.1007/s11280-012-0178-0.
31. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from Imbalanced Data in Surveillance of Nosocomial Infection. Artif Intell Med. 2006;37(1):7–18. https://doi.org/10.1016/j.artmed.2005.03.002.
32. Sahare M, Gupta H. A Review of Multi-class Classification for Imbalanced Data. Int J Adv Comput Res. 2012;2(3):160–4.
33. Gosain A, Sardana S. Handling Class Imbalance Problem Using Oversampling Techniques: A Review. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE; 2017. p. 79–85. https://doi.org/10.1109/icacci.2017.8125820.
34. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from Imbalanced Data Sets. Springer. 2018. https://doi.org/10.1007/978-3-319-98074-4.
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
36. Batista GEAPA, Prati RC, Monard MC. Balancing Strategies and Class Overlapping. In: Famili AF, Kok JN, Peña JM, Siebes A, Feelders A, editors. Advances in Intelligent Data Analysis VI. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 24–35.
37. Kosolwattana T, Liu C, Hu R, Han S, Chen H, Lin Y. A Self-inspected Adaptive SMOTE Algorithm (SASMOTE) for Highly Imbalanced Data Classification in Healthcare. BioData Min. 2023;16:15. https://doi.org/10.1186/s13040-023-00330-4.
38. Senge R, del Coz JJ, Hüllermeier E. On the Problem of Error Propagation in Classifier Chains for Multi-label Classification. In: Spiliopoulou M, Schmidt-Thieme L, Janning R, editors. Data Analysis, Machine Learning and Knowledge Discovery. Springer International Publishing; 2014. p. 163–170.
39. Liu B, Tsoumakas G. Dealing with Class Imbalance in Classifier Chains via Random Undersampling. Knowl-Based Syst. 2020;192:105292. https://doi.org/10.1016/j.knosys.2019.105292.
40. Charte F, Rivera A, del Jesus MJ, Herrera F. A First Approach to Deal with Imbalance in Multi-label Datasets. In: Pan JS, Polycarpou MM, Woźniak M, de Carvalho ACPLF, Quintián H, Corchado E, editors. Hybrid Artificial Intelligent Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 150–160.
41. Shamsudin H, Yusof UK, Jayalakshmi A, Akmal Khalid MN. Combining Oversampling and Undersampling Techniques for Imbalanced Classification: A Comparative Study Using Credit Card Fraudulent Transaction Dataset. In: 2020 IEEE 16th International Conference on Control & Automation (ICCA). IEEE; 2020. p. 803–808. https://doi.org/10.1109/icca51439.2020.9264517.
42. Perrot S, Bouhassira D, Fermanian J. Development and Validation of the Fibromyalgia Rapid Screening Tool (FiRST). Pain. 2010;150(2):250–6. https://doi.org/10.1016/j.pain.2010.03.034.
43. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with Pain Rating Scales. Ann Rheum Dis. 1978;37(4):378–81. https://doi.org/10.1136/ard.37.4.378.
44. Yates F. Contingency Tables Involving Small Numbers and the χ^2 Test. J R Stat Soc Ser B Stat Methodol. 1934;1(2):217–35. https://doi.org/10.2307/2983604.
45. Bonferroni CE. Il Calcolo delle Assicurazioni su Gruppi di Teste. In: Studi in onore del professore salvatore ortu carboni. 1935. p. 13–60.
46. Marchi H, Schmiegel S, Schamberger T, Fuchs C. Validating Methods for Inferring Co-occurring Diseases: A Flexible Framework for Simulating Synthetic Data. medRxiv. 2026. https://doi.org/10.64898/2026.01.13.26344018.
47. Kotelnikov A, Baranchuk D, Rubachev I, Babenko A. TabDDPM: Modelling Tabular Data with Diffusion Models. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning, vol. 202 of Proceedings of Machine Learning Research. PMLR; 2023. p. 17564–17579. https://proceedings.mlr.press/v202/kotelnikov23a.html.
48. R Core Team. R: A Language and Environment for Statistical Computing. Vienna; 2024. https://www.R-project.org/.
49. Python Software Foundation. Python. https://www.python.org/.

Publisher's Note