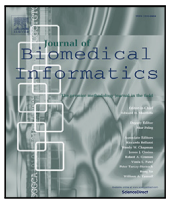
Journal of Biomedical Informatics 166 (2025) 104817

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

A novel machine learning-based workflow to capture intra-patient heterogeneity through transcriptional multi-label characterization and clinically relevant classification

Silvia Cascianelli *, Iva Milojkovic, Marco Masseroli

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milano, 20133, Italy

ARTICLE INFO

Keywords:
Multi-label classification
Trascriptional subtyping
Molecular heterogeneity
Clinically relevant stratification

ABSTRACT

Objectives: Patient classification into specific molecular subtypes is paramount in biomedical research and clinical practice to face complex, heterogeneous diseases. Existing methods, especially for gene expression-based cancer subtyping, often simplify patient molecular portraits, neglecting the potential co-occurrence of traits from multiple subtypes. Yet, recognizing intra-sample heterogeneity is essential for more precise patient characterization and improved personalized treatments.

Methods: We developed a novel computational workflow, named MULTI-STAR, which addresses current limitations and provides tailored solutions for reliable multi-label patient subtyping. MULTI-STAR uses state-of-the-art subtyping methods to obtain promising machine learning-based multi-label classifiers, leveraging gene expression profiles. It modifies standard single-label similarity-based techniques to obtain multi-label patient characterizations. Then, it employs these characterizations to train single-sample predictors using different multi-label strategies and find the best-performing classifiers.

Results: MULTI-STAR classifiers offer advanced multi-label recognition of all the subtypes contributing to the molecular and clinical traits of a patient, also distinguishing the primary from the additional relevant secondary subtype(s). The efficacy was demonstrated by developing multi-label solutions for breast and colorectal cancer subtyping that outperform existing methods in terms of prognostic value, primarily for overall survival predictions, and ability to work on a single sample at a time, as required in clinical practice.

Conclusions: This work emphasizes the importance of moving to multi-label subtyping to capture all the molecular traits of individual patients, considering also previously overlooked secondary assignments and paving the way for improved clinical decision-making processes in diverse heterogeneous disease contexts. Indeed, MULTI-STAR novel, reproducible and generalizable approach provides comprehensive representations of patient inner heterogeneity and clinically relevant insights, contributing to precision medicine and personalized treatments.

1. Introduction

Gene expression-based subtyping has become a valuable resource in biomedical research and clinical practice, especially for oncology. It enables classifying patients into distinct molecular subtypes, providing insights into disease heterogeneity and suggesting the most suitable treatment strategies [1–4]. However, the most widespread methods for cancer patient stratification primarily rely on similarity-based strategies [5–8] to find the most prominent subtype of each patient. These strategies perform dataset-level analyses and normalize the entire dataset across samples by applying global normalization techniques, such as quantile normalization or z-score transformation. This prevents working on a single sample at a time, as required in clinical practice. Furthermore, these strategies do not use standardized normalizations, which can lead to suboptimal reproducibility of subsequent similarity evaluations [9].

Furthermore, in many cases, state-of-the-art subtyping strategies may oversimplify the intricate landscape due to the intra-tumor heterogeneity. Studies have indeed highlighted the importance of recognizing the heterogeneity not only of a given type of cancer but even of a single tumor sample of a patient [10–12]. This inner heterogeneity, as confirmed by recent research [13–16], requires a more refined and comprehensive approach for improving patient characterizations and better identifying groups with shared molecular peculiarities and distinctive clinical traits.

* Corresponding author.
E-mail address: silvia.cascianelli@polimi.it (S. Cascianelli).

https://doi.org/10.1016/j.jbi.2025.104817
Received 6 November 2024; Received in revised form 13 March 2025; Accepted 14 March 2025
Available online 9 April 2025

Thus, to recognize molecular heterogeneity without losing the benefits of patient classifications for clinical utility, new approaches should still rely on the existing molecular subtypes of a given tumor, but provide more advanced patient stratification based on such subtypes. Instead of assigning a patient simply to a single subtype, it is essential to consider the possible memberships to multiple subtypes and their contributions to the mixed molecular and clinical traits of that patient. Multi-label classification can be the key to unveiling the intra-heterogeneity of a tumor sample, capturing the complete portrait of each patient and understanding different characteristics derived from more than a single molecular subtype. This shift in perspective can enhance precision medicine and personalized treatments, merging the advantages of a comprehensive patient-centric investigation and of a clinically relevant stratification.

Although multi-label classification techniques are widely used in applications focused on text, image and signal categorization, including various biomedical research tasks, their adoption in omics studies remains limited [16–19]. Prominent applications of multi-label classification can be found in image-based tasks to show evidence of diseases, such as in radiomics and digital pathology (e.g.,[20,21]), or in categorization tasks for structured and unstructured clinical data, often combined with Natural Language Processing techniques, in the medical informatics field, (e.g., [22–24]). In contrast, the application of multi-label classification in omics research, particularly in gene expression-based subtyping, holds a significant research opportunity. To the best of our knowledge, despite its high potential, a multi-label perspective in this area remains largely unexplored.

State-of-the-art transcriptional subtyping methods of a given cancer, (e.g., [6,7,25]) working on wide datasets normalized across samples, typically adopt similarity-based approaches to compute correlation or distance measures for each sample relative to all known subtypes. However, only the most similar subtype for each sample is assigned, performing a single-label classification. Substantial modifications would be required to move from continuous similarities to proper multi-label classification, allowing for multiple subtype assignments for each sample. Yet, this would not solve the main limitation of being unable to analyze just one sample at a time. Conversely, machine learning-based classifiers, which have recently been adopted for re-engineering subtyping tasks (e.g., [9,26–28]), enable single-sample, reproducible analysis of cancer patient profile; yet, so far they have barely evolved into multi-label solutions [16,22,29]. In fact, a framework capable of providing reliable multi-label references, training multi-label supervised models and ensuring generalization capabilities for subtyping tasks is still missing. Such a framework would enable the advanced recognition of multi-label subtypes for any new sample under exam, fostering improved patient stratification.

Here, we present MULTI-STAR (MULTI-label SubTyping and Advanced Recognition), a novel reproducible and generalizable methodological approach, implemented as a workflow and designed to address shortcomings of current subtyping methods by capturing overlooked heterogeneity in diseased patients and providing machine learning-based classifiers for comprehensive single-sample multi-label subtyping. MULTI-STAR fills the gap between the absence of multi-label references and the need for accurate multi-label predictions for a given transcriptional subtyping task. Leveraging gene expression profiles and a state-of-the-art similarity-based technique, it first generates multi-label characterizations of patients. Then, it employs these multi-label characterizations to train single-sample predictors, which are eventually capable of accurately assigning the needed (none, one or multiple) relevant subtypes for any new patient at a time. These assignments are also categorized to provide, for each patient, an advanced recognition of the most prominent subtype (hereafter referred to as primary subtype) from the additional, one or more, relevant subtypes (hereafter globally referred to as secondary subtypes). Notably, using MULTI-STAR, we developed multi-label classifiers of demonstrated prognostic value for well-known breast and colorectal cancers, taken as application examples.

Statement of significance	
Problem or Issue:	Current state-of-the-art methods for transcriptional subtyping generally overlook intra-patient heterogeneity.
What is Already Known:	This heterogeneity, largely due to co-existing cells with diverse or hybrid phenotypes, could be accurately captured by associating each patient with multiple molecular subtypes, when needed.
What this Paper Adds:	This paper introduces MULTI-STAR, a novel computational workflow enabling reliable transcriptional multi-label patient stratification. MULTI-STAR provides multi-label subtyping strategies recognizing, for each patient, the primary subtype and any additional relevant secondary subtypes. These strategies overcome the clinical value of single-label approaches, better capturing intra-patient heterogeneity and prognostic implications in cancer. Also, the modular workflow structure supports flexible applications to various diseases, subtypes, and classification tasks.

2. Methods

The following subsections present state-of-the-art methods and other related works of this study concerning both patient stratification, particularly in cancer genomics, and multi-label classification methodologies. First, we introduce existing gene expression-based classification approaches that provide ground truth single-label subtyping. MULTI-STAR indeed uses these approaches to obtain ground truth primary assignments and then dissects and extends them to extract proper multi-label characterization for every single sample under study. Following, we present machine learning paradigms and methods that MULTI-STAR employs for single-sample multi-label classification.

2.1. Similarity-based methods for cancer subtyping

Cancer subtyping is one of the most relevant examples of crucial patient stratification. It not only elucidates the molecular characteristics of tumors but also provides invaluable insights into patients' expected clinical outcomes. Subtyping research primarily focuses on tracing gene expression patterns and defining methods to distinguish cancer molecular subtypes. Many subtype systems [5,7,8,25,26,30–32], mostly different for each cancer type, have also been recognized for their outstanding clinical utility, especially in prognosis prediction and therapeutic decision-making; among others, they include the BRCA intrinsic subtypes (Basal, Her2-enriched - Her2, Luminal A - LumA, Luminal B - LumB, Normal-like) [4,9,30,33] and, more recently, the five CRC intrinsic subtypes (from CRIS-A to CRIS-E) [7]. Here, for CRC, CRIS subtypes are preferred over the alternative Consensus Molecular Subtypes (CMS) [8], since CRIS classes reflect stable characteristics of CRC cancer cells independently from the contribution of the surrounding stroma [7,16].

The majority of state-of-the-art subtyping approaches use similarity-based methods, which leverage various correlation or distance metrics to categorize an entire patient cohort using its corresponding gene expression dataset. First, the dataset is normalized to match an expected data distribution; then, each patient/sample of the dataset is assigned with its most similar molecular subtype. The PAM50 (Prediction Analysis of Microarray 50) test [6] is a widely recognized centroid-based assay for dataset-level BRCA subtyping, while the CRIS-NTP (ColoRectal cancer Intrinsic Subtypes-Nearest Template Prediction) [7] has been more recently developed to assign each colorectal cancer tumor of a patient cohort with one of the five CRIS subtypes. Specifically, the

PAM50 test uses a panel of 50 genes and their Spearman correlation with five centroids, one for each BRCA subtype, to find the subtype with the highest correlation for each sample of the cohort. The CRIS-NTP classifier, instead, leverages Z-score normalized sample profiles from a cohort of patients and calculates cosine distances to the five CRIS class templates; then, for each sample of the cohort it returns the subtype that exhibits the minimum significant distance to the sample (ensuring statistical significance with a Benjamini–Hochberg false discovery rate < 0.2).

2.2. Machine learning multi-label classification strategies

Single-label classification aims to correctly assign only the appropriate class to each sample under exam. Conversely, multi-label classification enables the assignment of multiple classes to a sample when needed. Therefore, it is a particularly challenging task, as it requires determining both whether the sample needs one or more assignments to be described and which specific labels are suited. To move from single-label to multi-label classification, MULTI-STAR explores strategies of *problem-transformation* and *algorithm-adaptation* [34,35].

Problem-transformation subdivides the multi-label problem into distinct single-label sub-problems, uses single-label classifiers as base learners and combines collected results through different approaches. We employ diverse *problem-transformation* methods adopting Logistic Regression (LR), k-Nearest Neighbors (kNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Adaptive Boosting (AdaB), Light Gradient Boosting (LGB), and Extreme Gradient Boosting (XGB) as base learners. The Binary Relevance method [36] trains separately distinct binary one-vs-all classifiers to independently predict each specific label, ignoring any correlation among labels. A Classifier Chain [37] adopts a chain of binary one-vs-all classifiers that operate sequentially. Each classifier, except the first one, incorporates in its feature space also training samples' predictions coming from the preceding classifiers in the chain. This sequential increase of the feature space enables considering relationships among labels, but the final performance is dependent on the heuristically chosen label order. The Label Powerset [36] method employs multi-class classifiers that consider as separate targets all the possible combinations of multiple labels observed in the training data. While accounting for label correlations, its prediction capability is constrained to assigning label combinations seen in the training phase. Two ensemble approaches incorporating problem transformation strategies are also evaluated. The Ensemble of Classifier Chains (ECC) [37] uses multiple independent Classifier Chains, varying label orderings to mitigate dependency, and aggregates results through majority voting. The RAndom K-labELsets (RAKEL) [38], is based on the Label Powerset approach and selects random combinations of k labels as target sets to learn specific k-label classification tasks with individual base learners. Each learner returns probability scores for every class included in the corresponding k-label set; these scores are then summed, averaged and thresholded to obtain the final predictions. Notably, this method can predict label combinations not present in the training data, differently from the Label Powerset approach.

Furthermore, we evaluate established methods of *Algorithm-adaptation*, explicitly designed for multi-label classification. The multi-label k-Nearest Neighbors (ML-kNN) [39] modifies the well-known single-label k-Nearest Neighbors classifier and incorporates the MAP (maximum a posteriori) principle to ascertain the association probability of a sample with each label independently of others. The multi-label Decision Tree algorithm (ML-DT) [40] has a modified entropy loss and optimizes splits by selecting features that best separate instances for the entire label set. The multi-label Adaptive Resonance Associative Map (ARAM) neural network [41] uses the adaptive resonance mechanism to learn and recognize the patterns present in the data efficiently, even when involving multiple labels. While other deep learning methods have shown promising results in handling multi-label classification [20,21], they are primarily designed for higher amounts of image-based training data to prevent overfitting, instability and optimization issues. Conversely, traditional machine learning approaches offer interpretability and robustness when working with more limited training datasets. These are both key aspects for achieving reliable patient stratification and clinically valuable insights in omics-based precision medicine. Therefore, given our transcriptional data of input and the limitations in sample sizes compared to feature dimensionality, traditional machine learning approaches combined with effective data preprocessing and feature selection (see Supplementary Subsection S1.1) are strongly preferred over deep approaches in our multi-label subtyping study. Yet, deep approaches could be included more in future applications, particularly in the case of integration with additional data modalities such as imaging.

3. Proposed MULTI-STAR workflow

The machine learning-based MULTI-label SubTyping and Advanced Recognition (MULTI-STAR) workflow presented here is schematically illustrated in Fig. 1. It serves a dual purpose for any specific subtyping task at hand:

- Establish a reliable reference of *multi-label characterization* for the available samples, by leveraging and extending similarity-based subtyping method at the state-of-the-art;
- Optimize a *multi-label classification* solution for subtyping, evaluating several machine learning models and strategies to identify the most promising and valuable multi-label classifier.

As mentioned, predicting multi-label assignments is challenging, requiring the estimation of both the correct number of assignments and the specific labels for each case. Reference labels are essential for providing reliable target examples to train and test any supervised multi-label classifier. Thus, the learning problem becomes even harder to address when dealing with scarcity or absence of data already associated with multiple labels. To address this, the MULTI-STAR initial automated step of *multi-label characterization* focuses on the pivotal task of generating multi-label references for the task at hand, also providing an adaptation of the pre-existing similarity-based subtyping methods to the multi-label framework. Following, the MULTI-STAR *multi-label classification* step rigorously explores strategies and models to overcome the limitations inherent in the dataset-based approaches used for the *multi-label characterization*. This step enhances the accuracy and robustness of the subtyping result, ensuring the portability of the method to any single, potentially heterogeneous sample under exam, with advanced recognition of all its relevant subtypes. Notably, MULTI-STAR is designed as a modular 'meta'-workflow that allows for flexible integration of alternative learning algorithms, adaptation strategies, and optimization methods. Thus, current choices represent only an initial selection that can be seamlessly modified, e.g., considering different strategies in expert-curated/knowledge-based steps for determining the best-performing solutions, or extended, e.g., integrating Bayesian optimization [42] for hyperparameter tuning or other approaches for classification.

Given the complexity of multi-label subtyping for a single sample, especially in cancer and other heterogeneous disease applications, evaluating the suitability of each method should not only consider an exhaustive performance assessment with a wide set of metrics, but also the prognostic capabilities of the obtained multi-label classifications. Thus, in our application use cases, we further compare all the best-performing solutions with each other based on their ability to recognize clinically relevant assignments (both primary and secondary). Particularly, we focus our knowledge-driven inspections on those subtypes that are most closely linked to the prognosis for each disease. Indeed, recent studies have underscored the importance of comprehensive, multi-label transcriptional classification in improving

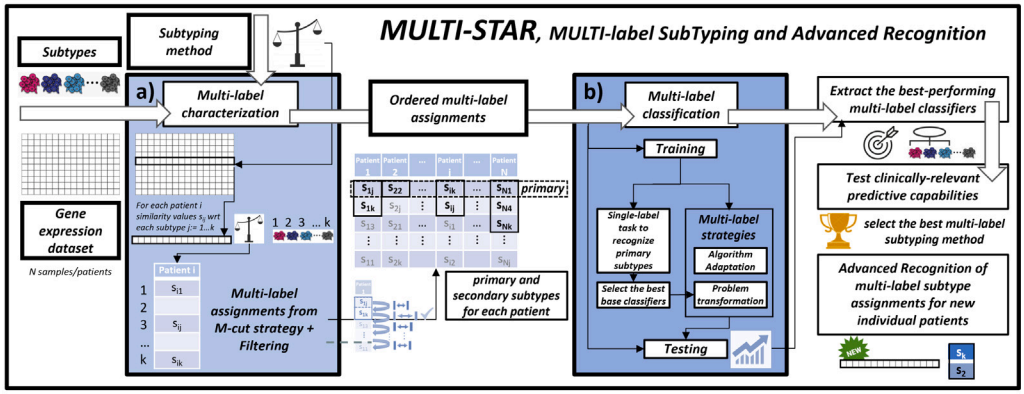


Fig. 1. Schematic description of the MULTI-STAR workflow with its two consecutive steps of: (a) multi-label characterization, to extract patient primary and secondary subtypes from expression data under study, and (b) multi-label classification, to obtain well-performing machine learning-based multi-label predictors of patients' subtypes.

predictions related to prognosis or response to treatments for cancer patients [15,16]. This emphasizes the potential impact of our MULTI-STAR workflow in advancing disease subtyping towards more precise personalized medicine.

3.1. MULTI-STAR multi-label characterization step

The first step of our MULTI-STAR workflow is generating a multi-label characterization for each sample, based on a state-of-the-art transcriptional subtyping approach and its stratification system. This step involves associating one or possibly multiple reference subtypes with each sample under study to capture a complete picture of the disease, even for inherently heterogeneous samples. Although most of the commonly used subtyping methods compute the similarities of any sample in a dataset with all the known subtypes, they do not provide any multi-label characterization and assign just the single, most prominent class to each sample of the dataset.

Conversely, MULTI-STAR generates multi-label references by leveraging similarity or dissimilarity measures from an existing state-of-the-art subtyping method. These can be, for example, correlations with subtype centroids, like in the PAM50 test [6] for BRCA, or cosine distances, like in the NTP-CRIS [7] for CRC samples. Notably, if the subtyping method uses a distance measure d, this latter can be easily transformed into its opposite (1-d) to obtain similarity values s between the instance under exam and each of the target subtypes of interest. Thus, for each sample we obtain a numerical vector with as many values of similarity as the number of target labels. In contrast to scenarios where high negative similarity (e.g., anti-correlation) carries significant meaning, MULTI-STAR approach focuses on the magnitude of similarity only in its positive direction. In fact, here negative similarity with respect to a given class does not provide relevant insights; it simply indicates greater distance from that class, potentially reflecting higher similarity to other classes, which are anyway evaluated independently. Thus, negative values are as non-informative as null values; reflecting the absence of positive similarity, they are here converted into zeros.

Then, the M-cut strategy [43] is applied to the numerical vector of each sample, ordered by decreasing similarity values, to derive one or multiple candidate subtypes. Indeed, M-cut pinpoints the largest difference between consecutive values in this ordered vector and partitions the corresponding subtype labels into two candidate groups: assignable labels with higher similarity and not assignable labels. Yet, the numerical vector is preserved to identify the most similar subtypes (i.e., primary subtype) and to better evaluate also the secondary subtypes among all the assignable labels. The M-cut strategy guarantees that each sample is assigned to its primary subtype, but it may inappropriately assign additional subtypes, even if the similarity values are all unusually low and marginal. To address this issue, we include a filter that excludes secondary labels with similarity values below a defined threshold. For each subtype, the threshold is computed as a percentile of the distribution of similarity values for the primary, hence steadier, subtype assignments in the training set. Exploring the effects of filtering at various percentiles, i.e., the 5th, 10th, and 25th, we have identified the 5th percentile to be the most suitable choice. This threshold effectively removes outliers within potential -secondary-label similarities without being overly restrictive compared to primary-label ones (see Supplementary Figures 1–2), and without significantly affecting the overall class distribution (see Supplementary Figures 3). Notably, this threshold can be appropriately fine-tuned using training data for any subtyping task.

Overall, the customized strategy combining M-cut and filtering is employed in the *multi-label characterization* step of our MULTI-STAR workflow to enhance the reliability of the multi-label assignments provided as references for the following *multi-label classification* step.

3.2. Multi-label classification step

The subsequent step of our MULTI-STAR workflow focuses on multi-label classification, aiming to find the most reliable machine-learning strategies for the given subtyping task. The accuracy metric guides a preliminary training and optimization process, examining all the base learners mentioned in Section 2.2. These models are initially assessed in a single-label context, predicting the primary subtype only (as detailed in Supplementary Subsection S3.1), to select the most suitable base learners for subsequent integration into multi-label problem transformation strategies (i.e., Binary Relevance, Classifier Chain, Label Powerset, ECC, and RAKEL). These latter strategies are compared with multi-label adapted algorithms (ML kNN, ML DT and ML ARAM) used as benchmarks.

For a meaningful comparison among all the classifiers, we split the data into stratified training and test sets using the 70:30 ratio and the same sample composition. This ensures a consistent subtype distribution in all sets, aligned with the single-label ground truth composition of the original dataset obtained from the existing gene expression-based classification approach, as described in Section 2. Every model optimization for multi-label classification is carried out using Grid Search for hyperparameter tuning and stratified 5-fold cross-validation to handle the complexity of multi-label assignments within a relatively restrained sample size. The suggested scoring metric for optimization is the weighted average F1 score. This can balance the contributions of all the subtype assignments differently from multi-label accuracy, which could be biased by major classes, or subset accuracy, which could be overly strict. Following, all the optimized classifiers

undergo thorough evaluation on the test set using an exhaustive set of metrics (see Section 3.3) to identify the most promising solutions for each classification task of interest.

3.3. Customized metrics for evaluating multi-label subtype predictions

In MULTI-STAR, a combination of newly defined measures and traditional label-based and example-based metrics [35,44–46] is employed to thoroughly evaluate the performance of each classifier in the complex and class-imbalanced multi-label setting of interest. Label-based metrics (here including precision, recall and F1 score) assess the independent local performance of each class considering only samples assigned with that class label; then micro, macro and weighted averaging are used to estimate global performances across the classes. Multi-label example-based metrics (here including subset accuracy, multi-label accuracy, average precision and Hamming loss), instead, assess the prediction performance on each sample separately and then average the obtained values; thus, they are precious to handle multiple labels of a sample equally and simultaneously. For further details, please refer to Supplementary Subsection S1.2. Moreover, when class labels display a relevant ordering, different example-based metrics can be used (e.g. the average precision) or even defined to provide accurate performance evaluations that directly consider label ranking.

In our multi-label classification, we are clearly interested in correctly assigning all the needed subtypes to depict a patient. Yet, in subtyping tasks, we also aim to distinguish the most prominent primary subtype from all the additional (one or more) secondary subtypes of a patient. Thus, the ranking of predicted labels of each multi-label classifier (derived from label assignment probabilities) is used to distinguish the primary subtype from other secondary assignments, if any, for every patient. Accordingly, to offer further indications not provided by traditional example-based metrics, we designed customized metrics that, for each sample, compare the ranking of the predicted labels against the ranking of the reference labels obtained within the MULTI-STAR characterization step.

- The *relaxed accuracy* evaluates the model's ability to assign the primary reference label while disregarding whether this label is designated as the primary or a secondary label in the predicted assignments.
- The *primary-ordered accuracy* requires the primary class in the reference assignments to be confirmed as the primary class among the predicted assignments. In contrast to the relaxed accuracy, which only needs the presence of the primary class among the predicted labels, this metric is more stringent.
- The *secondary-ordered accuracy* is similar to the primary-ordered accuracy but for each sample specifically assesses the confirmation of any secondary class in the reference assignments as secondary predictions in the predicted assignments.
- The *ordered subset accuracy*, which is the most stringent metric and requires complete overlap of label predictions and rankings between the predicted assignments and the reference ones.

To favor reader understanding, a graphical representation of these original evaluation metrics introduced in MULTI-STAR methodological approach is provided in Fig. 2.

3.4. Clinical validation

Prognostic implications of the stratification obtained from a multi-label classifier can be tested using clinical annotations, when available, to validate the clinical relevance. Specifically, overall survival or time to disease recurrence/progression annotations are insightful in assessing the prognostic relevance of the found multi-label predictions, as we demonstrate in our application use cases, discussed in the following. Given MULTI-STAR multi-label predictions involving a specific subtype *sub*, we can examine two different analytical scenarios and patient partitions: (1) primary *sub* patients vs. non-*sub* at all (i.e., neither primary nor secondary); (2) *sub* patients (both primary or secondary) vs. non-*sub* at all. Clinical events occurring in these patient partitions can be analyzed to test the prognostic value of the best-performing multi-label classifiers or compare them with each other and with their simpler single-label classification counterparts. Particularly, Kaplan–Meier curves are used to visualize differences between the groups, while log-rank tests estimate the statistical significance of such differences.

4. Application use cases

The effectiveness of the proposed MULTI-STAR workflow is demonstrated through its application to well-known subtyping problems for two extensively studied tumor types: breast and colorectal cancer. Specifically, we collected gene expression profiles [47], focusing on BRCA and CRC datasets from The Cancer Genome Atlas (TCGA) [48]. Please, for data preprocessing and feature selection [49] steps, refer to Supplementary Subsection S1.1.

4.1. MULTI-STAR application to breast cancer subtyping

To deal with the BRCA subtyping use case, for each sample of the corresponding B_TCGA dataset, we first extracted the Pearson correlation values of similarity to all the PAM50 [6] centroids of the BRCA intrinsic subtypes. These similarity values were processed as described in Section 3.1, using M-cut strategy and filtering techniques to obtain a reference multi-label characterization for all the patient samples. Corresponding primary and secondary class distributions are reported in Supplementary Figure 3a, where it is evident that Normal-like secondary assignments are extremely numerous. This occurs almost exclusively in combination with primary LumA assignments, which is not surprising; in fact, the Normal-like class refers to samples from largely unaffected tissue and the LumA class includes many cases showing better-expected prognoses [4,50]. All subsequent multi-label classifiers also confirmed this relationship.

Before training multi-label classifiers using this reference characterization as supervised information, we determined the most suitable learners for single-label BRCA subtyping to be employed as base models in our problem-transformation strategies (see Section 3.2 and Supplementary Subsection S3.1). By comparing the optimized models (in Supplementary Table 1), four of them (LR, SVM, RF and XGB) overcame the others. Their confusion matrices and local metrics of class-specific performances are reported in Supplementary Figure 4.

In the following multi-label classification step, the four selected base learners were exhaustively employed within all the different problem transformation techniques considered, in comparison to the alternative approaches of algorithm adaptation, detailed in Section 2.2. All multi-label classifiers were properly trained and optimized using the F1 score in 5-fold-cross-validation and the multi-label characterization provided by the first step of our MULTI-STAR workflow as supervised information. A wide range of multi-label example-based local and global metrics was adopted to adequately assess the performance of each classifier in subtyping test samples. However, none of the assessed algorithm adaptation solutions achieved overall performances surpassing problem transformation and ensemble approaches (see Supplementary Table 3). Performance metrics of the adapted models were worse than those highlighted in bold in Tables 1 and 2. Notably, all the newly defined ordered (subset, primary and secondary) accuracy metrics appeared extremely disappointing, showing the inadequacy in this context of the adapted models in the absence of ad-hoc adaptations, i.e., multi-label adaptation strategies that are not general-purpose but specifically tailored for the task of interest as in [16].

As we can observe from the values in Table 1, LR and XGB models mostly outperformed the others regarding average precision and both subset and multi-label accuracies, also when considering our

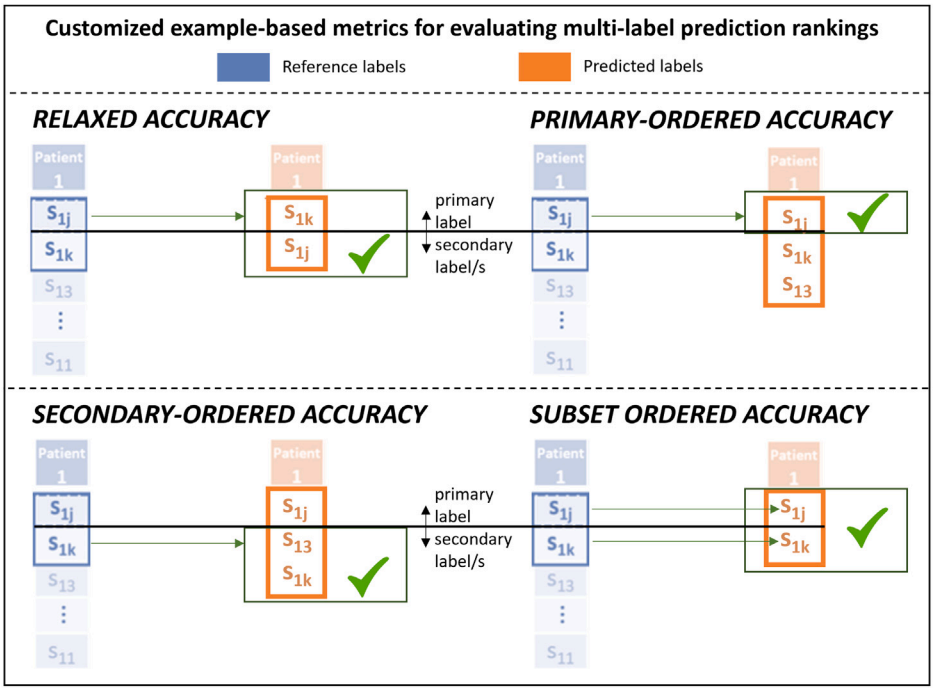


Fig. 2. Graphical representation of the four multi-label evaluation metrics newly defined. They compare the ranking of the reference labels obtained by the MULTI-STAR characterization step with the ranking of the predicted labels, derived from the label assignment probabilities of each of the MULTI-STAR predictors obtained in the MULTI-STAR classification step. For each sample i, s_{ij} is its score of the association with the class j, according to the reference MULTI-STAR characterization (in blue) or the predictions of the MULTI-STAR classifier under exam (in orange). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Global performance measures on BRCA subtyping of each optimized multi-label strategy of problem transformation from evaluations of standard and new metrics. Bold values indicate metrics outperforming other problem-transformation solutions and all the adapted models.

		Subset accuracy	Multi-label accuracy	Relaxed accuracy	Average precision	Hamming loss	Ordered subset accuracy	Ordered accuracy	
								Primary	Secondary
XGB	**Binary Relevance**	**0.772**	**0.872**	0.949	**0.961**	0.051	0.680	0.839	0.699
	Classifier Chain	0.763	**0.865**	0.946	0.957	0.054	0.699	0.864	0.728
	Label Powerset	0.763	**0.866**	0.937	0.935	0.056	**0.718**	0.877	0.744
	ECC	**0.772**	**0.869**	0.949	0.946	0.051	0.684	0.839	0.712
	RAKEL	**0.782**	**0.876**	**0.956**	**0.962**	**0.048**	**0.725**	0.873	**0.753**
Logistic Regression	**Binary Relevance**	**0.772**	**0.867**	0.934	**0.962**	**0.050**	**0.718**	**0.880**	**0.747**
	Classifier Chain	**0.772**	**0.867**	0.946	**0.962**	**0.050**	**0.718**	**0.880**	**0.747**
	Label Powerset	0.756	**0.871**	**0.962**	0.956	0.056	0.693	**0.889**	0.709
	ECC	**0.772**	0.863	0.946	0.956	0.051	**0.718**	**0.880**	**0.756**
	RAKEL	0.737	0.851	**0.953**	0.957	0.058	0.680	0.873	0.718
SVM	Binary Relevance	0.737	0.855	**0.953**	0.958	0.057	0.684	**0.892**	0.715
	Classifier Chain	0.737	0.855	**0.953**	0.958	0.057	0.684	**0.892**	0.715
	Label Powerset	**0.772**	**0.874**	**0.956**	**0.966**	0.054	0.709	0.873	0.731
	ECC	0.741	0.852	**0.953**	0.947	0.059	0.687	**0.892**	0.718
	RAKEL	0.753	0.862	**0.959**	0.958	0.056	0.639	0.832	0.668
Random Forest	Binary Relevance	0.766	0.863	0.934	**0.960**	0.053	0.693	0.845	0.725
	Classifier Chain	0.759	0.860	0.937	0.958	0.054	0.680	0.842	0.715
	Label Powerset	0.759	0.864	0.930	0.945	0.058	0.693	0.839	0.718
	ECC	0.766	0.859	0.934	0.952	0.054	0.693	0.845	0.734
	RAKEL	0.750	0.849	0.924	0.942	0.059	0.665	0.820	0.715

customized ordered metrics. Furthermore, these two models exhibited high and stable global performances when using several averaging methods over the class-specific local measures, as indicated in Table 2.

Particularly, LR brought remarkable performances across all the different multi-label strategies, showing robustness across all classes and stable balancing between all pairs of averaged precision and recall. Conversely, the XGB reached its best results when combined with Binary Relevance or ensemble approaches (i.e., ECC and RAKEL) and tended to favor precision across all classes, limiting their false positive rates. Among all the other models, the SVM used within a Label Powerset strategy returned encouraging results and appeared worthy of further investigation. Accordingly, LR, XGB and SVM models, each one in combination with its most suitable multi-label strategies, were further investigated and compared with prognosis-based assessments to verify the clinical value of their provided stratification, as described in Section 4.3.

4.2. MULTI-STAR application to colorectal cancer subtyping

To address our CRC subtyping application, for each sample of the corresponding C_TCGA dataset, we extracted all cosine distances

Table 2
Global performance measures on BRCA subtyping of each optimized multi-label strategy of problem transformation obtained by averaging subtype-specific metrics. Bold values indicate metrics outperforming other problem-transformation solutions and all the adapted models.

		Macro			Micro			Weighted		
		Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
XGB	**Binary Relevance**	**0.941**	0.857	**0.892**	**0.932**	**0.897**	**0.914**	**0.935**	**0.897**	**0.913**
	Classifier Chain	0.917	0.840	0.872	0.926	0.891	0.908	0.926	0.891	0.905
	Label Powerset	**0.930**	0.804	0.858	**0.937**	0.872	0.903	**0.936**	0.872	0.899
	ECC	**0.932**	0.865	**0.892**	0.921	**0.903**	**0.912**	0.924	**0.903**	**0.911**
	RAKEL	**0.933**	0.862	**0.892**	0.928	**0.908**	**0.918**	0.929	**0.908**	**0.916**
Logistic Regression	**Binary Relevance**	0.924	**0.888**	**0.906**	**0.932**	**0.899**	**0.915**	**0.932**	**0.899**	**0.915**
	Classifier Chain	0.924	**0.888**	**0.906**	**0.932**	**0.899**	**0.915**	**0.932**	**0.899**	**0.915**
	Label Powerset	0.894	0.870	0.881	0.916	**0.895**	0.905	0.916	**0.895**	0.905
	ECC	0.911	**0.893**	**0.902**	0.921	**0.906**	**0.913**	0.921	**0.906**	**0.913**
	RAKEL	0.878	**0.888**	0.883	0.903	**0.901**	0.902	0.905	**0.901**	0.903
SVM	Binary Relevance	0.895	0.871	0.883	0.918	0.891	0.904	0.917	0.891	0.904
	Classifier Chain	0.895	0.871	0.883	0.918	0.891	0.904	0.917	0.891	0.904
	Label Powerset	0.903	0.859	0.880	**0.930**	0.888	0.909	0.929	0.888	0.908
	ECC	0.880	0.872	0.876	0.905	**0.895**	0.900	0.905	**0.895**	0.899
	RAKEL	0.886	**0.883**	0.885	0.907	0.901	0.904	0.907	**0.901**	0.904
Random Forest	Binary Relevance	**0.952**	0.808	0.859	**0.941**	0.878	0.908	**0.945**	0.878	0.903
	Classifier Chain	**0.937**	0.809	0.858	**0.941**	0.876	0.907	**0.942**	0.876	0.903
	Label Powerset	0.926	0.802	0.853	**0.930**	0.874	0.901	**0.930**	0.874	0.897
	ECC	**0.943**	0.814	0.859	**0.930**	0.884	0.906	**0.933**	0.884	0.902
	RAKEL	0.912	0.824	0.861	0.915	0.882	0.898	0.915	0.882	0.895

with respect to each of the five subtype templates of the CRIS-NTP algorithm [7]. Notice that for CRIS-NTP, the primary label corresponds to the subtype that has reached the smallest cosine distance to the sample under study. Thus, we subtracted each cosine distance from 1, to obtain similarity values that enabled us to apply exactly the same workflow as for BRCA. The similarity values of each CRC sample underwent the M-cut strategy and filtering (as detailed in Section 3.1), to obtain a reference multi-label characterization used as supervised information for the subsequent classification step. The distribution of primary and secondary assignments of CRC patients for each CRIS class is summarized in Supplementary Figure 3b.

As for the BRCA use case, before training multi-label classifiers, we identified the most suitable learners to use as base models in problem-transformation strategies (see Section 3.2 and Supplementary Subsection S3.1). Among such models, optimized for CRC subtyping, LR, SVM, RF, and XGB outperform the others (see Supplementary Table 2). Supplementary Figure 5 includes the confusion matrices and the histograms of local metrics for these four models to illustrate class-specific performances. The selected base learners were used within all the different problem transformation and ensemble techniques considered in the multi-label classification step and compared with each other and the multi-label adapted algorithms. As for BRCA subtyping application, we abandoned the adapted algorithms due to their underperformance (see Supplementary Table 3), particularly serious for the newly defined ordered accuracies, which again highlighted the need for ad-hoc adaptations to effectively follow this alternative paradigm for improving CRC subtyping [16].

From the values in Table 3, SVM and LR models emerge as the best-performing, especially considering the multi-label and newly defined primary-ordered accuracy. These two models brought remarkable performances across different multi-label strategies, showing particular robustness with Binary Relevance and Ensemble of Classifier Chains across global measures from overall evaluations and from averaging methods, as shown in Table 4. In addition, the XGB model combined with the Label Powerset strategy reached the highest subset accuracy and comparable multi-label and primary-ordered accuracies and was therefore considered worthy of further investigation. Instead, the RF model was discarded due to the imbalance between all the pairs of averaged precision and recall. Eventually, SVM and LR, each combined with all its most appropriate multi-label strategies, and XGB with Label Powerset were selected to verify the clinical value of their provided stratifications, as described in the following Section 4.3, and to identify the most clinically relevant approach for multi-label subtyping of CRC patients.

4.3. Clinical value assessments

For clinical applications, MULTI-STAR classifiers must not only represent the molecular heterogeneity of a sample in terms of its multi-label assignments, but also provide reliable estimates of prognosis, treatment response, or other clinically relevant implications based on the predicted subtypes. Thus, here we show how to employ the clinical outcome annotations available for our datasets to identify the most promising multi-label classifiers for advanced recognition of the correct subtype(s) of each patient, also considering their prognostic capabilities. The outcomes of these assessments also demonstrate the clinical relevance of multi-label predictions provided by MULTI-STAR.

Given the well-established associations of specific subtypes to expected prognosis for both BRCA and CRC, a careful study of clinical events and patient partitions of interest was conducted (see also Supplementary subsection S3.3). For every considered subtype *sub*, both primary and secondary assignments from each multi-label classifier of interest were examined. Consequently, every non-*sub* group included only patients not assigned to subtype *sub* neither as a primary nor secondary class. In addition, for each model type, the results of a single-label classifier trained to recognize the primary subtype only were reported for further comparison. This comparison was crucial to explore any prognostic improvements brought by multi-label classification over a classical, less informative, single-label stratification.

4.3.1. Prognostic assessments on BRCA multi-label stratifications

From performance evaluations in the context of BRCA, LR models combined with any multi-label strategy, XGB when using Binary Relevance or ensemble methods, and SVM with a Label Powerset approach appeared worthy of further investigations to assess the clinical value of their provided stratification. We focused on the LumA subtype, known for its better long-term prognosis compared to other BRCA subtypes, and we grouped patients in primary LumA, primary or secondary LumA, and NOT-LumA at all. Then, we evaluated two events for each of these patient groups: survival after 10 years and disease recurrence within 5 years. Based on the predictions of each model, in Table 5, we report the p-value (statistical significance for p-value <= 0.05) of

Table 3
Global performance measures on CRC subtyping of each optimized multi-label strategy of problem transformation from evaluations of standard and new metrics. Bold values indicate metrics outperforming other problem-transformation solutions and all the adapted models.

		Subset accuracy	Multi-label accuracy	Relaxed accuracy	Average precision	Hamming loss	Ordered subset accuracy	Ordered accuracy	
								Primary	Secondary
XGB	Binary Relevance	0.543	0.674	0.763	0.865	0.117	0.532	0.720	0.710
	Classifier Chain	0.565	0.674	0.758	0.866	0.118	0.548	0.720	0.720
	Label Powerset	**0.667**	**0.759**	0.817	0.848	0.105	**0.629**	0.774	0.747
	ECC	0.548	0.677	0.763	0.859	0.113	0.538	0.720	0.720
	RAKEL	0.602	0.708	0.769	0.860	0.110	0.591	0.747	**0.763**
Logistic Regression	**Binary Relevance**	0.634	**0.762**	**0.866**	0.882	**0.092**	0.591	**0.790**	0.694
	Classifier Chain	0.602	0.740	0.855	0.875	0.096	0.559	0.769	0.677
	Label Powerset	**0.640**	**0.763**	0.849	0.880	0.103	**0.629**	**0.812**	0.720
	ECC	**0.640**	**0.764**	**0.866**	**0.895**	**0.088**	0.597	**0.790**	0.704
	RAKEL	0.624	**0.757**	0.839	0.883	0.101	0.605	**0.796**	0.715
SVM	**Binary Relevance**	0.608	**0.763**	**0.892**	0.879	**0.092**	0.570	**0.790**	0.651
	Classifier Chain	0.608	**0.763**	**0.892**	0.890	**0.092**	0.570	**0.790**	0.651
	Label Powerset	0.634	**0.756**	0.823	**0.887**	0.104	**0.618**	0.780	0.720
	ECC	0.613	**0.766**	**0.892**	**0.904**	**0.090**	0.575	**0.790**	0.656
	RAKEL	0.629	0.746	0.833	0.881	0.102	0.597	0.758	0.737
Random Forest	Binary Relevance	0.527	0.631	0.694	0.847	0.118	0.511	0.667	0.742
	Classifier Chain	0.538	0.631	0.688	0.851	0.116	0.516	0.640	0.758
	Label Powerset	0.634	0.724	0.763	0.855	0.117	**0.629**	0.758	**0.763**
	ECC	0.527	0.631	0.694	0.855	0.114	0.511	0.667	0.758
	RAKEL	0.570	0.674	0.720	0.848	0.111	0.565	0.704	**0.769**

Table 4
Global performance measures on CRC subtyping of each optimized multi-label strategy of problem transformation obtained by averaging subtype-specific metrics. Bold values indicate metrics outperforming other problem-transformation solutions and all the adapted models.

		Macro			Micro			Weighted		
		Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
XGB	Binary Relevance	0.801	0.672	0.723	0.806	0.698	0.748	0.809	0.698	0.742
	Classifier Chain	0.790	0.676	0.724	0.802	0.698	0.747	0.799	0.698	0.741
	Label Powerset	**0.843**	0.694	0.757	**0.842**	0.711	0.771	**0.843**	0.711	0.767
	ECC	0.801	0.684	0.730	0.806	0.711	0.755	0.810	0.711	0.750
	RAKEL	0.825	0.676	0.733	0.821	0.706	0.759	0.823	0.706	0.749
Logistic Regression	**Binary Relevance**	0.806	0.792	**0.798**	0.820	0.806	**0.813**	0.823	0.806	**0.814**
	Classifier Chain	0.810	0.777	0.792	0.821	0.789	0.804	0.826	0.789	0.805
	Label Powerset	0.809	0.734	0.767	0.821	0.750	0.784	0.821	0.750	0.781
	ECC	0.806	0.807	**0.805**	0.820	**0.820**	0.820	0.824	0.820	**0.821**
	RAKEL	0.794	0.773	0.782	0.799	0.785	0.792	0.799	0.785	0.791
SVM	**Binary Relevance**	0.794	**0.812**	**0.800**	0.807	**0.828**	**0.817**	0.811	**0.828**	**0.817**
	Classifier Chain	0.794	**0.812**	**0.800**	0.807	**0.828**	**0.817**	0.811	**0.828**	**0.817**
	Label Powerset	0.808	0.734	0.764	0.817	0.7520	0.782	0.820	0.750	0.779
	ECC	0.791	**0.824**	**0.804**	0.803	**0.838**	**0.820**	0.808	**0.838**	**0.820**
	RAKEL	0.791	0.764	0.774	0.801	0.776	0.788	0.806	0.776	0.788
Random Forest	Binary Relevance	**0.844**	0.606	0.691	**0.851**	0.638	0.792	**0.848**	0.638	0.714
	Classifier Chain	**0.871**	0.595	0.693	**0.873**	0.625	0.729	**0.870**	0.625	0.713
	Label Powerset	**0.845**	0.630	0.697	0.825	0.672	0.741	0.839	0.672	0.721
	ECC	**0.844**	0.616	0.699	**0.851**	0.649	0.736	**0.848**	0.649	0.721
	RAKEL	**0.871**	0.598	0.685	**0.870**	0.645	0.741	**0.871**	0.645	0.716

any log-rank test comparing the survival curves of a pair of patient groups, as indicated in the table header. Multi-label classifiers are indicated in the first column, while corresponding single-label models for primary class prediction are reported in the additional fourth and seventh columns, one for each event.

As we can observe from the p-values, the two events under study show a clear difference in the contribution given by the secondary LumA assignments across all the assessed models. The expected better clinical outcome emerged in terms of overall survival after 10 years as a peculiar trait of all Luminal A patients, both primary and secondary. This is highlighted by the significant p-values associated with the primary LumA assignments and strongly confirmed (even with higher significance in some models) when also including secondary LumA cases with respect to the NOT-LumA samples at all. Conversely, the lower risk of recurrence known to be associated with LumA samples here was mostly confirmed as a strong indicator for Primary LumA samples. These latter mostly denote a statistically significant lower rate of recurrences within 5 years compared to NOT-LumA cases, while this significance weakened when also considering secondary LumA assignments.

Thus, both these clinical events can be better predicted for BRCA patients when using a more complete multi-label stratification. Indeed, the contribution of secondary underlying assignments appeared relevant to better capture key differences, while none of the cases of simpler patient partitions based on single-label subtyping employing the same learning algorithm reached adequate significance. This clearly demonstrates the value of advanced multi-label recognition. Notably, it highlights the non-negligible effects of considering so-far hidden secondary assignments to obtain more molecularly accurate and prognostically relevant patient stratifications, which may help improve clinical handling and therapeutic decision processes.

Table 5
Statistical significance of the log-rank tests comparing the survival curves of different pairs of BRCA patient groups when considering each of the two clinical events reported on the top of the table. Significant p-values (<= 0.05) are highlighted in bold, as well as the classifiers bringing the most clinically relevant multi-label stratifications in recognizing LumA samples with better-expected prognosis.

Multi-label classifiers	Patient decease within 10 years			Tumor recurrence within 5 years		
	Primary LumA vs. NOT-LumA	Any LumA vs. NOT-LumA	Single-label LumA vs. all the others	Primary LumA vs. NOT-LumA	Any LumA vs. NOT-LumA	Single-label LumA vs. all the others
XGB with Binary Relevance	**0.0143**	**0.0130**	XGB: 0.0922	0.122	0.0794	XGB: 0.0824
XGB with Ensemble CC	**0.0143**	**0.0130**		0.122	0.0794	
XGB with RAKEL	**0.00792**	**0.0125**		**0.0213**	0.0794	
LR with Binary Relevance	**0.0153**	**0.0150**	LR: 0.102	**0.0253**	0.0674	LR: 0.0829
LR with Classifier Chain	**0.0153**	**0.0150**		**0.0253**	0.0674	
LR with Label Powerset	**0.0216**	**0.0168**		**0.0414**	0.0956	
LR with Ensemble CC	**0.0153**	**0.0150**		**0.0253**	0.0674	
LR with RAKEL	**0.0132**	**0.0170**		**0.0282**	0.0956	
SVM with Label Powerset	**0.0320**	**0.0139**	SVM: 0.114	0.0838	**0.0449**	SVM: 0.0780

Besides this key evidence, arising from the vast majority of the assessed models, the results in Table 5 allowed us to select the most promising multi-label classifier(s) for BRCA subtyping, also considering their prognostic capabilities. Accordingly, we selected the XGB combined with the RAKEL approach and, among the prognostically equivalent options (i.e., LR with Binary Relevance, Classifier Chain, or Ensemble of Classifier Chains), we selected the LR with Classifier Chain, which was slightly more robust on all the previous computational performance evaluations (see Section 4.1). Kaplan-Meyer curves based on the assignments of the XGB with RAKEL and LR with Classifier Chain models, for both survival after 10 years and recurrence within 5 years events, are depicted in Supplementary Figures 6–7. While the LR was more balanced in global metrics derived from class-specific evaluations (see Table 2), the XGB reached better results in overall performance measures from global evaluations (see Table 1). Both options appeared sound and valid, with the first slightly more tailored for sensitivity and the second for precision. Therefore, both these solutions can perform valuable advanced recognition of BRCA multi-label subtypes on single-patient gene expression profiles, while their joint use could even further strengthen the reliability of the obtained predictions.

4.3.2. Prognostic assessments on CRC multi-label stratifications

In the context of colorectal cancer, we concentrated the clinical evaluation on the CRIS-B subtype, which has the poorest expected prognosis compared to the other subtypes; accordingly, for CRC patients we assessed the decease event within a 5-year time horizon. The compared models were the best ones on the previous computational performance evaluations on multi-label CRC subtyping (see Section 4.2). Considering the multi-label predictions of every model, Table 6 presents the statistical significance derived from log-rank tests when comparing the survival curves of each pair of patient groups specified in the table header, among primary CRIS-B, primary or secondary CRIS-B, and NOT-CRIS-B at all. As we can notice, there is a confirmed strong association between a bad prognosis and CRIS-B patients, as almost all the stratifications have significant results from log-rank tests considering the standard threshold of 0.05. This clinical trait appeared to be stronger for primary assignments despite being conserved also when including secondary assignments. The comprehensive multi-label stratification of CRC patients demonstrates superior predictability of the decease clinical event compared to single-label subtyping for all SVM-based models and for many of the other approaches.

From the results in Table 6, we identified the XGB combined with the Label Powerset approach as the most promising multi-label classifier for CRC subtyping when considering the crucial predictive role of the CRIS-B class, including samples with the worst expected clinical outcomes. Despite this capability being confirmed by the vast majority of the assessed classifiers, the XGB with Label Powerset demonstrates an enhanced prognostic power considering multi-label assignments compared both to the other models and to its single-label counterpart. Kaplan-Meyer curves based on its assignments and examining the decease within 5 years event are reported in Supplementary Figure 8.

5. Discussion and conclusions

We introduced our innovative MULTI-STAR computational approach for multi-label transcriptional subtyping, designed to comprehensively recognize heterogeneity in gene expression profiles of diseased patients and assign one or more molecular subtypes. We demonstrated that MULTI-STAR overcomes the shortcomings of single-label state-of-the-art subtyping methods, which often have a dataset-level implementation, suboptimal reproducibility and miss crucial aspects of disease heterogeneity. Conversely, MULTI-STAR provides reliable, single-sample predictors of established subtypes, able to capture the heterogeneity of disease biology more effectively through multi-label classification, and whose results are fully reproducible on the same data due to its standardized steps and precise parametrizations. Indeed, MULTI-STAR offers a well-defined methodological framework ensuring analytical reproducibility and flexibility through its modular workflow implementation.

A key strength of MULTI-STAR lies in its ability to modify any existing similarity-based approach and subtyping standard towards multi-label characterization. However, this also represents a current limitation, as MULTI-STAR requires established methods and target subtypes to be straightforwardly applied to any cancer or heterogeneous disease. Nonetheless, its modular design can ease future extensibility, including potential integration with unsupervised clustering or with classifications based on other multi-omics or biomedical input data, like histological, image or functional data.

In the considered application use cases, the *multi-label characterization* step of MULTI-STAR guaranteed sound and consistent references for multi-label subtyping of both BRCA and CRC patients, considering as targets their well-known intrinsic subtypes. Then, MULTI-STAR predictors exhibited effectiveness in their advanced recognition of the appropriate multi-label assignments for individual BRCA and CRC samples. Particularly, BRCA and CRC subtyping tasks highlighted the versatility and reliability of problem transformation strategies without the need for ad-hoc modifications to obtain tailored multi-label adapted strategies. Indeed, in both use cases, general-purpose adaptation techniques were underperforming, likely also due to their reliance on single classifiers rather than strategies combining multiple classifiers, as in problem transformation approaches.

Table 6
Statistical significance of the log-rank tests comparing the survival curves of different pairs of CRC patient groups when considering the decease within 5 years as the clinical event of interest. Significant p-values (<= 0.05) are highlighted in bold, while the asterisk marks significant p-values even smaller than 0.005. The multi-label classifier bringing the most clinically relevant stratification in recognizing CRIS-B samples with worse expected prognosis is highlighted in bold.

Multi-label classifiers	Patient decease within 5 years		
	Primary CRIS-B vs. NOT-CRIS-B	Any CRIS-B vs. NOT-CRIS-B	Single-label CRIS-B vs. all the others
XGB with Label Powerset	**0.0012***	**0.0038***	XGB: **0.0329**
LR with Binary Relevance	**0.0475**	0.0626	
LR with Label Powerset	**0.0192**	**0.0222**	LR: **0.0289**
LR with Ensemble CC	**0.0475**	0.0626	
SVM with Binary Relevance	0.0063	**0.0105**	
SVM with Classifier Chain	0.0063	**0.0105**	SVM: **0.0118**
SVM with Label Powerset	**0.0025***	**0.0134**	
SVM with Ensemble CC	0.0063	**0.0105**	

Collected results for BRCA and CRC demonstrated that primary assignments alone can be inadequate to fully characterize a sample, showing the relevance of both primary and secondary assignments to offer more precise indications concerning specific clinical events. These clinical validations stressed the noteworthy role of the so-far overlooked secondary assignments in obtaining more molecularly accurate and prognostically relevant patient stratifications. Nonetheless, they also confirmed the importance of specifying the most prominent primary subtype compared to meaningful but secondary assignments. In this context, we clearly demonstrated also the relevance of the new example-based multi-label metrics that we introduced in our MULTI-STAR methodological approach: indeed, these metrics enhance performance evaluation by considering prediction rankings and distinguishing between primary and secondary label categorizations.

Overall, our innovative MULTI-STAR workflow for multi-label patient characterization and classifications can bridge the gap between subtyping research and translational medicine for different tumors and other heterogeneous diseases. Additionally, beyond its current implementation for transcriptional subtyping, future extensions could open its use to other biomedical or multi-omics contexts in the face of evolving research.

In conclusion, the presence of multiple subtypes co-existing within a single patient is evidently the result of a broader spectrum of molecular traits. These can be identified through a multi-label framework, in contrast to the limited focus on the predominant subtype, which has been conventionally considered so far at the state-of-the-art. The shift towards a multi-label perspective of each patient, especially in oncogenomics, has the potential to significantly advance precision medicine, improving clinical outcome predictions and personalized treatment options. To this aim, the versatility and effectiveness of MULTI-STAR is a noteworthy contribution, as already demonstrated by its successful application in enhancing the molecular and prognostic value of BRCA and CRC patient stratifications.

CRediT authorship contribution statement

Silvia Cascianelli: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Iva Milojkovic:** Validation, Software, Methodology. **Marco Masseroli:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Professor Enzo Medico for the fruitful discussions on the topic.

Appendix A. Supplementary data

The developed Python code implementing the MULTI-STAR workflow is publicly available at https://github.com/DEIB-GECO/MULTI-STAR.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2025.104817.

References

[1] A. Schlicker, G. Beran, C.M. Chresta, G. McWalter, A. Pritchard, S. Weston, S. Runswick, S. Davenport, K. Heathcote, D.A. Castro, et al., Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines, BMC Med. Genom. 5 (1) (2012) 1–15.
[2] M. Ringnér, G. Jönsson, J. Staaf, Prognostic and chemotherapy predictive value of gene-expression phenotypes in primary lung adenocarcinoma, Clin. Cancer Res. 22 (1) (2016) 218–229.
[3] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, B. Shi, Breast cancer intrinsic subtype classification, clinical use and future trends, Am. J. Cancer Res. 5 (10) (2015) 2929–2943.
[4] J. Holm, L. Eriksson, A. Ploner, M. Eriksson, M. Rantalainen, J. Li, P. Hall, K. Czene, Assessment of breast cancer risk factors reveals subtype heterogeneity, Cancer Res. 77 (13) (2017) 3708–3717.
[5] S. Biade, M. Marinucci, J. Schick, D. Roberts, G. Workman, E. Sage, P. O'Dwyer, V. Livolsi, S. Johnson, Gene expression profiling of human ovarian tumours, Br. J. Cancer 95 (8) (2006) 1092–1100.
[6] J.S. Parker, M. Mullins, M.C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (2009) 1–10.
[7] C. Isella, F. Brundu, S.E. Bellomo, F. Galimi, E. Zanella, R. Porporato, C. Petti, A. Fiori, F. Orzan, R. Senetta, et al., Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer, Nat. Commun. 8 (2017) 1–16.
[8] J. Guinney, R. Dienstmann, X. Wang, A. De Reynies, A. Schlicker, C. Soneson, L. Marisa, P. Roepman, G. Nyamundanda, P. Angelino, et al., The consensus molecular subtypes of colorectal cancer, Nature Med. 21 (11) (2015) 1350–1356.
[9] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, E. Medico, Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer, Sci. Rep. 10 (1) (2020) 1–13.
[10] S. Ogino, C.S. Fuchs, E. Giovannucci, How many molecular subtypes? Implications of the unique tumor principle in personalized medicine, Expert. Rev. Mol. Diagn. 12 (6) (2012) 621–628.
[11] A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D.P. Cahill, B.V. Nahed, W.T. Curry, R.L. Martuza, et al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science 344 (6190) (2014) 1396–1401.
[12] Q. Wang, B. Hu, X. Hu, H. Kim, M. Squatrito, L. Scarpace, A.C. DeCarvalho, S. Lyu, P. Li, Y. Li, et al., Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment, Cancer Cell 32 (1) (2017) 42–56.

[13] S. Ma, S. Ogino, P. Parsana, R. Nishihara, Z. Qian, J. Shen, K. Mima, Y. Masugi, Y. Cao, J.A. Nowak, et al., Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis, Genome Biol. 19 (1) (2018) 1–14.
[14] F.A. Büttner, S. Winter, V. Stühler, S. Rausch, J. Hennenlotter, S. Füssel, S. Zastrow, M. Meinhardt, M. Toma, C. Jerónimo, et al., A novel molecular signature identifies mixed subtypes in renal cell carcinoma with poor prognosis and independent response to immunotherapy, Genome Med. 14 (1) (2022) 1–19.
[15] L. Marisa, Y. Blum, J. Taieb, M. Ayadi, C. Pilati, K. Le Malicot, C. Lepage, R. Salazar, D. Aust, A. Duval, et al., Intratumor CMS heterogeneity impacts patient prognosis in localized colon cancer, Clin. Cancer Res. 27 (17) (2021) 4768–4780.
[16] S. Cascianelli, C. Barbera, A.A. Ulla, E. Grassi, B. Lupo, D. Pasini, A. Bertotti, L. Trusolino, E. Medico, C. Isella, et al., Multi-label transcriptional classification of colorectal cancer reflects tumor cell population heterogeneity, Genome Med. 15 (1) (2023) 1–37.
[17] E.A. Tanaka, S.R. Nozawa, A.A. Macedo, J.A. Baranauskas, A multi-label approach using binary relevance and decision trees applied to functional genomics, J. Biomed. Inform. 54 (2015) 85–95.
[18] L. Xie, S. He, Y. Wen, X. Bo, Z. Zhang, Discovery of novel therapeutic properties of drugs from transcriptional responses based on multi-label classification, Sci. Rep. 7 (1) (2017) 7136.
[19] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, O. Schwengers, D. Heider, Multi-label classification for multi-drug resistance prediction of escherichia coli, Comput. Struct. Biotechnol. J. 20 (2022) 1264–1270.
[20] M. Irtaza, A. Ali, M. Gulzar, A. Wali, Multi-label classification of lung diseases using deep learning, IEEE Access (2024).
[21] H. Lai, Q. Yao, Z. He, X. Tao, S.K. Zhou, Long-tailed multi-label classification with noisy label of thoracic diseases from chest X-ray, in: 2024 IEEE International Symposium on Biomedical Imaging, ISBI, IEEE, 2024, pp. 1–5.
[22] Z. Ceylan, E. Pekel, Comparison of multi-label classification methods for prediagnosis of cervical cancer, Graph Model. 21 (2017) 22.
[23] Y. Guo, F.-L. Chung, G. Li, L. Zhang, Multi-label bioinformatics data classification with ensemble embedded feature selection, IEEE Access 7 (2019) 103863–103875.
[24] Q. Chen, J. Du, A. Allot, Z. Lu, LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (5) (2022) 2584–2595.
[25] M.D. Wilkerson, X. Yin, V. Walter, N. Zhao, C.R. Cabanski, M.C. Hayward, C.R. Miller, M.A. Socinski, A.M. Parsons, L.B. Thorne, et al., Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation, PLoS One 7 (5) (2012) 1–13.
[26] A. Kamoun, A. de Reyniès, Y. Allory, G. Sjödahl, A.G. Robertson, R. Seiler, K.A. Hoadley, C.S. Groeneveld, H. Al-Ahmadie, W. Choi, et al., A consensus molecular classification of muscle-invasive bladder cancer, Eur. Urol. 77 (4) (2020) 420–433.
[27] F. Cristovao, S. Cascianelli, A. Canakoglu, et al., Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (1) (2020) 121–134.
[28] S. Mongardi, S. Cascianelli, M. Masseroli, Biologically weighted LASSO: enhancing functional interpretability in gene expression data analysis, Bioinformatics 40 (10) (2024) btae605.
[29] A.A. Alyousef, S. Nihtyanova, C.P. Denton, P. Bosoni, R. Bellazzi, A. Tucker, Latent class multi-label classification to identify subclasses of disease for improved prediction, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2019, pp. 535–538.
[30] T. Sørlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. Van De Rijn, S.S. Jeffrey, et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc. Natl. Acad. Sci. 98 (19) (2001) 10869–10874.
[31] P.S. Mischel, R. Shai, T. Shi, S. Horvath, K.V. Lu, G. Choe, D. Seligson, T.J. Kremen, A. Palotie, L.M. Liau, et al., Identification of molecular subtypes of glioblastoma by gene expression profiling, Oncogene 22 (15) (2003) 2361–2373.
[32] D.J. McConkey, W. Choi, Molecular subtypes of bladder cancer, Curr. Oncol. Rep. 20 (2018) 1–7.
[33] The cancer genome atlas network, comprehensive molecular portraits of human breast tumours, Nature 490 (7418) (2012) 61–70.
[34] J.M. Nareshpalsingh, H.N. Modi, Multi-label classification methods: A comparative study, Int. Res. J. Eng. Technol. (IRJET) 4 (12) (2017) 263–270.
[35] S. DMongardi, M. Masseroli, S. Cascianelli, Supervised learning: Multi-label classification, in: Reference Module in Life Sciences, Elsevier, 2024.
[36] M.R. Boutell, J. Luo, X. Shen, C.M. Brown, Learning multi-label scene classification, Pattern Recognit. 37 (2004) 1757–1771.
[37] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (2011) 333–359.
[38] G. Tsoumakas, I. Katakis, I. Vlahavas, Random k-labelsets for multilabel classification, IEEE Trans. Knowl. Data Eng. 23 (2011) 1079–1089.
[39] M.L. Zhang, Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognit. 40 (2007) 2038–2048.
[40] A. Clare, R.D. King, Knowledge discovery in multi-label phenotype data, in: L.D. Raedt, A. Siebes (Eds.), Proceededings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery: PKDD'01, Springer, Berlin, 2001, pp. 42–53, http://dx.doi.org/10.1007/3-540-44794-6_4.
[41] A. Tan, Adaptive resonance associative map, Neural Netw. 8 (3) (1995) 437–446.
[42] V. Nguyen, Bayesian optimization for accelerating hyper-parameter tuning, in: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering, AIKE, IEEE, 2019, pp. 302–305.
[43] C. Largeron, C. Moulin, M. Géry, MCut: A thresholding strategy for multi-label classification, in: International Symposium on Intelligent Data Analysis, Springer, 2012, pp. 172–183.
[44] G. Tsoumakas, I. Katakis, Multi-label classification: an overview, Int. J. Data Warehous. Min. 3 (2007) 1–13.
[45] R.B. Pereira, A. Plastino, B. Zadrozny, L.H. Merschmann, Correlation analysis of performance measures for multi-label classification, Inf. Process. Manage. 54 (3) (2018) 359–369.
[46] S. Mongardi, S. Cascianelli, M. Masseroli, Performance measures for multi-class classification, in: Reference Module in Life Sciences, Elsevier, 2024.
[47] S. Pallotta, S. Cascianelli, M. Masseroli, RGMQL: scalable and interoperable computing of heterogeneous omics big data and metadata in R/Bioconductor, BMC Bioinformatics 23 (1) (2022) 123.
[48] J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R.M. Shaw, B.A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J.M. Stuart, The Cancer Genome Atlas pan-cancer analysis project, Nature Genet. 45 (10) (2013) 1113–1120.
[49] S. Cascianelli, A. Galzerano, M. Masseroli, Supervised relevance-redundancy assessments for feature selection in omics-based classification scenarios, J. Biomed. Inform. 144 (2023) 1–12.
[50] O. Yersal, S. Barutca, Biological subtypes of breast cancer: Prognostic and therapeutic implications, World J. Clin. Oncol. 5 (3) (2014) 412–426.