Received: May 27, 2023. Revised: June 11, 2023. 

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# An Effective Random Forest Approach for Mining Graded Multi-Label Data

**Wissal Farsal[1]*** **Mohammed Ramdani[1]** **Samir Anter[1]**

[1]*Department of Computer Science, Faculty of Sciences and Technics of Mohammedia, Hassan II University of Casablanca, Morocco*
* Corresponding author's Email: farsalwissal@gmail.com

**Abstract:** The graded multi-label classification (GMLC) is an extension of multi-label classification. Whilst a multi-label classifier is limited to predicting the set of relevant labels, a graded multi-label classifier predicts the degree of relevance of a set of given labels. A key challenge of this learning problem consists of modelling the dependencies among the labels to improve the predictive accuracy. The algorithm adaptation-based solutions, which modify the algorithms directly to handle GMLC were proven effective in modeling these dependencies in comparison to the transformation-based models which reduce the graded multi-label datasets into a set of multi-class or binary datasets. In this paper, we propose an adaptation of random forest algorithm (GML_DT) with an adapted CART (classification and regression trees). The adapted algorithm is based on a modified formula for the Gini Index which fully models the label dependencies. The performance of the new model was tested on a 101 benchmark datasets and compared against the most influential methods for GMLC. The evaluation metrics considered in the experimental study are proper to the graded multi-label setting, i.e., hamming loss and vertical 0-1 loss. The experimental results show that the proposed model outperforms the considered models for the graded multi-label performance metrics. The increase in performance and overall accuracy is quantified by a decrease in the hamming loss of more than 13% and a decrease in vertical 0-1 loss that exceeds 10% when compared to the state-of-the-art models.

**Keywords:** Graded multi-label classifiers, Adaptation algorithms, CART, Random forest, Graded multi-label decision trees.

## 1. Introduction

Multi-label classification (MLC) tackles the problem of learning where each instance is associated with multiple labels simultaneously [1-3]. MLC was successfully applied in different fields as it provides a solution to overcome the limitations of standard classification methods [4]. An example of such applications is the text categorization problem which is the task of predicting the different categories to which a text belongs. This is prompted by the fact that documents usually cover more than one topic at once. Examples of implementation of MLC methods on different types of documents include topic classification in news articles [5, 6], classification of medical files, web documents, etc. [7]. Another key application of MLC is in multimedia content annotation such as images, sound, and videos [8, 9]. Furthermore, MLC played an important role in bioinformatics, and disease diagnostic. Especially with gene function prediction and protein function prediction [10, 11].

However, in most real-world applications, the relevance of the labels is rather ordinal and varies in its level for each label. By constricting a classification to a crisp binary format, the multi-label construct remains inefficient as to the information it yields. The graded multi-label classification (GMLC) was introduced as an extension of MLC to address this issue. The graded multi-label learning allows the labels to have membership degrees and therefore constitutes a more informative and fitting alternative to MLC.

GMLC extends the multi-label setting by assigning a degree of membership on an ordinal scale for each class label. The relevance is therefore represented by a fuzzy set instead of the crisp binary

set {0, 1}. The membership set is comprised of a finite subset of contiguous integers that can be considered as linguistic descriptors, e.g., a high-octane, fairly suspenseful serious movie would be represented by a set of memberships {'5', '3', '1'} on an ordinal scale of one to five corresponding to the categories 'Action', 'Suspense', and 'Comedy'. Hence, graded multi-label classification provides more information as to how relevant the labels are for a given data sample. For a streaming service, for example, this additional information can be highly useful to their clients.

This added value can be utilized in a variety of other real-world applications. Recently, in [12], the graded multi-label setting was demonstrated to be a more fitting paradigm for Music Emotion Recognition than the standard single-label and multi-label approaches. Furthermore, GMLC was successfully deployed as a framework for recommendation systems [13-15].

In addition to the movies dataset, which categorizes the genres of the movies based on their degrees of relevance, another functional use of GMLC is explored in this article. That is mining data from surveys and questionnaires, where it is common to have answers on a graded scale to quantify questions of importance, quality, or agreement i.e. it is arguably easier for people to express their level of agreement with a question on a graded scale, meanwhile, they may hesitate to give a definitive 'yes' or 'no' answer.

Additionally, by simply considering the available multi-label datasets, one can notice that many of these datasets would be far more useful and informative with the inclusion of the membership degrees of the labels, e.g., the datasets IMDB, emotions and reuters [16] etc.

As an extension of the multi-label learning, a graded multi-label classifier can be applied to a MLC problem. However, the reverse is not true, GMLC tasks require new methods that factors in the fuzzy memberships or relevance of the labels.

Several different approaches were developed to solve the graded multi-label classification task. These methods can be categorized under two main paradigms, namely, the transformation paradigm and the adaptation paradigm.

On the one hand, the former relies on a simple key idea of reducing the GMLC problem into a combination of standard multi-class or binary classifications and thus use the standard existing methods to solve them. A straightforward example of a problem transformation would be to reduce the problem into a number of independent and ordinal classifiers, one for each class label, otherwise known as binary relevance (BR) [17]. Evidently, this solution is easy to implement through a simple conversion of the graded multi-label dataset into several multi-class ones (one for each label) and therefore standard multi-class algorithms can be applied on the resulting datasets. Albeit its simplicity, BR completely ignores the underlying dependencies between the labels which lead to a loss in information and a decline in accuracy. Different transformation-based approaches were proposed with the aim of taking into account these interdependencies [13-15, 18-20]. These methods relied on intuitive transformation schemas, thus decomposing the original problem into a number of multi-label classification tasks that are then solved using various approaches based on RPC (ranking by pairwise comparison), CLR (calibrated label ranking), IBLR (instance based logistic regression), etc. [16]. However, these methods were either limited to partially modelling the label dependencies or unable to correctly learn them.

On the other hand, the adaptation methods directly modify the classification algorithms to handle the graded multi-label data. In [21], an adapted decision tree was proposed. This new model modified the entropy function of the original algorithm to extend and generalize it to handle GMLC. Thus resulting in a single tree that fully models label interdependencies and predicts the degrees of all labels simultaneously.

Moreover, this algorithm was applied in the research field of music emotion recognition (MER) and was shown to be highly effective in uncovering and understanding the underlying features that are directly associated with a specific set of levels for different perceived emotions in music [12]. This uncovered information is instrumental in this field of research. Therefore proving the merits of the adaptation-based paradigm over the transformation-based one.

However, to the extent of our knowledge, this solution remains the only adaptation-based approach for GMLC. Thus, we focused our attention to developing a new adapted method that improve the current existing one. Considering that the proposed decision tree classifier is prone to overfitting, there is a clear need for a more robust model that is also capable of handling graded multi-label data.

In this article, we propose an adapted graded multi-label Random Forest in combination with an adapted CART algorithm.

The main contributions of this paper are stated as follows:

- We propose a new formula for Gini Index,

one that generalized the heuristic to measure the divergences between the probability distributions and the target membership degrees of all the labels.

- We developed a graded multi-label CART algorithm (GML_CART) based on the new heuristic. The GML_CART is capable of fully modelling the label interdependencies. Additionally, it has the advantage of interpretability. This model produces a single tree capable of predicting a vector corresponding to the membership degrees of a given set of labels. Thus generating comprehensible and meaningful rules.
- We then developed a random forest algorithm based on the adapted CART. This model combines the effectiveness of the modified CART in handling GMLC and the robustness of the random forest classifier that stems from the bagging technique (bootstrap aggregation) as well as the random subspace method.
- Finally, we assessed the performance of the new models using the evaluation metrics that fit the graded multi-label setting such as the hamming loss to calculate the divergence of the predictions and the vertical 0-1 loss which is a generalized accuracy metric. We conducted and experimental study on 101 datasets to compare the new models to the most prevalent graded multi-label classifiers from the literature. Our evaluation results show that GML_RF outperformed the existing models in terms of hamming loss and vertical 0-1 Loss.

The rest of the paper is organized as follows. In section 2, we review graded multi-label classifiers. We introduce the GML-RF algorithm and the GML_CART algorithm in section 3. We then display the results of the conducted experiments on graded multi-label data in section 4. Finally, we conclude this work in section 5.

## 2. Related work

In GMLC, we can distinguish between two fundamental approaches: adaptation and transformation. The former aims to efficiently adapt classifiers to the graded multi-label setting, whereas in the latter, the original problem of learning is broken into a number of separate classifiers, e.g., a combination of independent classifiers, one for each label. Thus, the transformation is applied on the datasets in order to convert them into several multi-class or multi-label datasets. Because they can utilize existing algorithms, the initial solutions that were proposed for GMLC were transformation-based.

In [20], the authors introduced two main transformation schemata for reducing the graded multi-label classification into individual multi-class tasks or separate multi-label problems.

The vertical reduction approach [20] transforms the task of GMLC into several independent ordinal classification tasks, by simply considering each label as a separate single label classification problem. This basic transformation method can be considered as the equivalent of binary relevance (BR) in multi-label learning. Where it decomposes a multi-label classification problem into independent binary classification problems (one per each label). Albeit their simplicity, these methods are based on the assumption of independence between the labels. Which is a major shortcoming that results in the loss of information and a lower accuracy [19, 17].

The second transformation schema for GMLC is called the horizontal reduction [20]. It operates by decomposing the graded multi-label learning task into a combination of multi-label learning tasks. This technique trains a classifier for each degree of membership and predicts the labels that are associated with the given degree. Consequently, it is capable of partially modelling the label dependencies unlike the previous method that completely ignores them. However, the horizontal reduction fails to comply with the trivial monotony rule that exists in the data, i.e., if a label is associated with a certain membership degree then it is also associated with all the membership degrees that are less or equal to the given degree. Furthermore, this flaw leads to contradictory predictions [19].

In [20], the authors used the horizontal reduction with IBLR-ML a multi-label classifier combining instance-based learning with logistical regression in order to take into account the label interdependencies. However, this method showed worse results than the baseline binary relevance [19].

In [19], the authors proposed three methods based on an extension of calibrated label ranking (CLR) [16] in MLC. The working principle of these extensions is similar to that of CLR, the only distinction lies in the use of a set of virtual labels instead of one. The first method, namely, the Horizontal_CLR, consists of using the aforementioned horizontal decomposition and solving the ensuing multi-label problems with the CLR transformation approach. This method not only inherits the drawback related to the horizontal reduction in causing contradictory predictions, but the pairwise classifications also result in a discrepancy between the generated binary classifiers

for the virtual labels and the generated ones for the rest of the labels. Thus, causing a bias in the predictions.

The second method in [19] is called full-CLR. It provides an alternative that extends the Horizontal-CLR and overcomes some of its shortcomings. However, it remains unable to exploit the information about the difference between the degrees of membership of the class labels.

Therefore, a third approach, namely joined-CLR was proposed in order to overcome these limitations to a certain degree by combining both methods. This approach generalizes the aforementioned methods by producing classifiers for joint ranking across all degrees of membership and labels including the virtual labels. Thus balancing the number of preferences between the labels. Although this is expected to solve the problem of biased predictions in Horizontal_CLR, the added preferences about the virtual labels are trivial and constitute a bias on which the classifications are found.

Furthermore, all these methods are not able to model the dependencies between the labels since they are limited to learning the pairwise label preferences.

In [18], the authors addressed this issue by introducing an approach for learning label dependencies and label preferences. The proposed transformation approach combined CLR and pairwise comparisons (RPC) and PSI (pre-selection, selection and interest of chaining) which allows for learning the label dependencies. However, this approach was outperformed by the previous methods and failed to learn the correct label dependencies.

In [21], a novel adapted graded multi-label decision tree classifier (GML_DT) was introduced. This adaptation was mainly achieved by generalizing the formula of the entropy to cover GMLC tasks. While the standard entropy for single-label classification quantifies the purity of a set according to the target values, the adapted entropy measures the purity of a set according to multiple labels with their respective degrees of membership.

The reasoning behind the proposed adaptation stems from the fact that the standard entropy deriving from Shannon's entropy in information theory is in a unit of bits. Thus allowing for one bit per class, whilst the new formula extents the information described by the entropy to model the memberships of the class labels.

Using this adapted heuristic, GML_DT was shown to be very competitive with state-of-the-art approaches [21]. Furthermore, it is the only interpretable method capable of producing a single decision tree that predicts a set of degrees associated with the label set at its leaves. Which in turn can be converted into a set of simple and useful rules.

However, GML_DT can be prone to overfitting. Hence the need for a more robust model. Therefore, we propose a graded multi label random forest that models the label dependencies and prevents overfitting by using multiple randomized trees.

## 3. The proposed methods

In a graded multi-label dataset, an instance $\boldsymbol{x}$ is represented by a vector of d elements $\boldsymbol{x} = (x_1, \dots, x_d)$ drawn from a d-dimensional input space $\mathbb{X} = X_1, \dots, X_d$ of numerical or categorical attributes. Each instance is associated with a set of predefined labels $L = \{\lambda_1, \dots, \lambda_k\}$ to certain degrees of relevance. These degrees of relevance or membership form a predefined set of ordered values $M = \{\mu_1, \dots, \mu_m\}$ where $\mu_1 < \mu_2 < \cdots < \mu_m$. For an instance $\boldsymbol{x}$ we map each class label $\lambda \in L$ to its degree of membership, ranging from $\mu_1$ which represents the complete irrelevance of a label to $\mu_m$ which means the full membership of the instance to the label in question. With this description, we can see that GMLC is a generalization of multi-label learning classification where only two degrees are permitted $M = \{0, 1\}$.

Given a training set of N graded samples $S = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N}$ where $\boldsymbol{y}^{(i)} = (y_1^{(i)}, \dots, y_k^{(i)})$ is the vector of degrees of membership associated with the instance $\boldsymbol{x}^{(i)}$ and where $y_j^{(i)} \epsilon M$ represents the degree of membership corresponding to the label $\lambda_j$, $j = 1, \dots, k$. We aim to build a classifier $H$ that predicts the degrees of memberships of each label for a new instance $\boldsymbol{x}$.

$$H: \mathbb{X} \longrightarrow M^k \qquad \boldsymbol{x} \mapsto \boldsymbol{y} \tag{1}$$

The output of this classifier for a new instance $\boldsymbol{x}$ is the prediction vector $\hat{\boldsymbol{y}}$ defined as follows:

$$H(\boldsymbol{x}) = \hat{\boldsymbol{y}} = [\hat{y}_1, .., \hat{y}_k] \tag{2}$$

where $\hat{y}_i$, $i = 1, \dots, k$ is the predicted membership degree of $\boldsymbol{x}$ in regards to the $i$th class label $\lambda_i$.

### 3.1 Graded multi-label CART algorithm

The process of classification with the CART algorithm relies on learning a binary tree by recursively partitioning the input space according to the Gini index. Which is a widely used heuristic for constructing decision trees.

The Gini index is an impurity-based criterion

measuring the divergences between the probability distributions of the target attribute values [22].

For a data set S, the standard Gini is defined as:

$$Gini(S) = 1 - \sum_{j=1}^{n} p_j^2 \quad (3)$$

where $n$ is the number of classes and $p_j$ is the probability of the $j$th class.

In order to handle graded multi-label problems, we need to modify the formula for multiple class labels. The standard Gini index calculates the misclassification of a single class. Therefore, we propose a way to generalize this formula to account for the misclassifications of the class labels by calculating the Gini index as the average sum of the label-wise Gini indices.

The graded multi-label Gini index is calculated as follows:

$$GML\_Gini(S) = \frac{1}{k}\sum_{i=1}^{k} Gini(S, \lambda_i) \quad (4)$$

$Gini(S, \lambda_i)$ is the label-wise Gini index calculated for the sample set S while only considering the label $\lambda_i$, it is defined as:

$$GML\_Gini(S, \lambda_i) = 1 - \sum_{j=1}^{m} p_{\mu_j}^2 \quad (5)$$

where $p_{\mu_j}$ is the probability of the $j$th grade $\mu_j$.

This allows the heuristic to account for the overall impurity of a partition regarding all class labels simultaneously.

**Algorithm 1. Pseudocode for GML_CART**

Given a training set D with d attributes
Create a node N containing D
if stopping criterion is True
    exit
else
1. For each possible attribute-value split (A,v):
   Calculate the total Gini Index induced by splitting the node on (A,v):
   $$Gini_{split}(D) = \sum_{i\in\{1,2\}} \frac{|D_i|}{|D|} Gini(D_i)$$
2. Split the node into two children nodes based on the test (A,v) that minimizes the Gini impurity
3. Recursively apply GML_CART on each child node

GML_CART, the adapted graded multi-label CART described in Algorithm 1, follows the same induction approach for building decision trees as the standard CART algorithm. Thus, the classification tree is constructed top-down in a greedy manner. For each node, the algorithm searches for the best attribute-value test for partitioning the remaining training samples. The best attribute-value test is chosen by considering all possible split values or points for each attribute and selecting the one with the smallest Gini index. After splitting the samples in the node to create two children nodes, one for which the test succeeds and one for which the test fails, the algorithm calls itself recursively on each child node until a stopping criterion is met.

Table 1. A preview of the toy dataset

| Instances | $a_1$ | $a_2$ | $a_3$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ |
|---|---|---|---|---|---|---|
| $x_0$ | A | 51 | 9 | 3 | 0 | 0 |
| $x_1$ | C | 27 | 11 | 3 | 3 | 1 |
| $x_2$ | B | 13 | 17 | 1 | 0 | 2 |
| $x_3$ | C | 35 | 15 | 2 | 1 | 0 |

The stopping conditions for the developed model are the number of samples remaining in a node is less than a predefined threshold. The second condition is the samples belong to the same degree class per label. The third criterion concerns the depth of the tree i.e. the construction of the tree stops when a maximum predefined depth is reached.

When a stopping criterion is met, GML_CART creates a leaf node containing the vector of predictions relative to the relevance of the labels. The predicted vector is obtained via majority vote to determine the most frequent degree of membership for each label.

To illustrate the functionality and the value of using GML_CART, we apply it to a toy dataset. Table 1 represents a subset of this dataset which contains 37 samples. Each sample is described by 3 attributes and is associated with a set of labels $\{\lambda_1, \lambda_2, \lambda_3\}$ to some degree on a predefined scale of 4 contiguous values $\{0, 1, 2, 3\}$.

Fig. 1 shows the decision tree constructed using the GML_CART algorithm. This decision tree is binary, where each internal node represents an attribute-value test and the two resulting branches are the two possible outcomes i.e. the test is true or the test is false.

The best split tests are chosen based on the overall weighted GML_Gini induced by the split on the attribute-value. This measure chooses the splits that reduce the impurity of the partitions with respect to all the class labels.

The added value of this tree is in the leaf nodes which allow for the prediction of the degrees of the labels simultaneously. The GML_CART tree

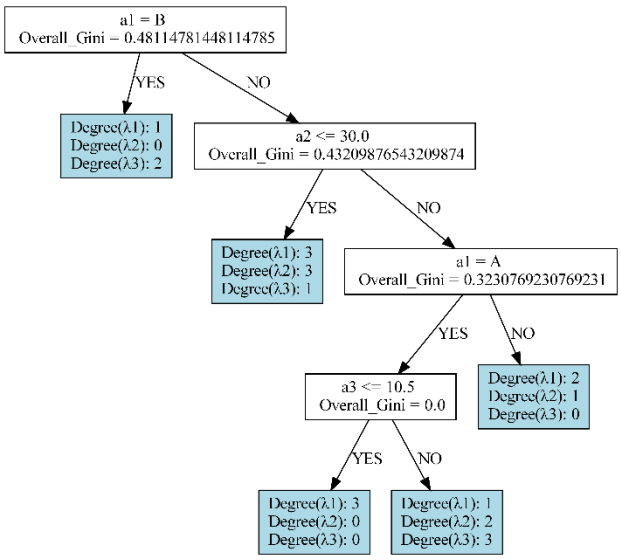


Figure. 1 Example of a graded multi-label CART decision tree

considers the intrinsic label dependencies in the graded multi-label dataset and determines the relevant attribute-value conditions for the prediction of the grades of the three labels which can also be written as simple and informative decision rules.

### 3.2 Graded multi-label random forest

Random forest algorithm figures amongst the most popular and successful decision tree-based ensembles. The purpose of this type of classifier is to combine many weak learners into a stronger one [23].

The algorithm relies on a simple modus operandi: construct many randomized decision trees or CART-based trees to classify a new instance using the majority vote. The randomized trees are built using bagging (bootstrap aggregation) which generates bootstrap samples from the dataset and aggregates the learners by majority vote.

Bagging is an ensemble method that constructs each classifier on a different sample from the dataset. This is achieved by re-sampling with replacement from the original dataset. This process is known as bootstrapping and the created subset is called "inbag". The "inbag" is the same size as the training set and contains duplicate samples. The bootstrapping technique usually leaves out one-third of the training set and the set of excluded examples is referred to as "out of bag" (OOB).

The bootstrap aggregation process reduces the variance for algorithms with high variance such as decision trees. In the case of CART, the algorithm tends to overfit the data which results in lower prediction performance. Whereas using bagging in combination with CART leads to a more stable and accurate model.

Additionally, the decision nodes are partitioned using the CART-split criterion on a randomly selected subset drawn from the set of features. The randomization increases the diversity of the generated trees which leads to higher predictive performance when their results are combined.

The graded multi-label random forest (GML_RF) we propose follows Breimen's methodology. During the learning phase, the algorithm constructs 100 trees without pruning. Each tree is constructed using a different bootstrap sample in the same manner as bagging. In the classification phase, the algorithm uses the generated trees to classify a new instance by a majority vote. Algorithm 2 defines the basic outline to generate the graded multi-label ensemble.

**Algorithm 2. Pseudocode for GML_RF**

Training phase:
Given a training set D
To generate C CART classifiers:
for i=1 to C do
1. Randomly sample the training data D with replacement to produce Di
2. Call GML_CART(Di)

end for
Classification phase:
- Combine the decisions of the C CART classifiers using a label-wise majority vote.

**Algorithm 3. Pseudocode for the modified GML_CART**

Given a training set D with d attributes
Create a node N containing D
if N is pure then
exit
else
1. Randomly select $\sqrt{d}$ of the splitting features in N
2. Find the attribute-value split that minimizes the GML_Gini Index and use it to split the node into two children nodes
3. Recursively apply the GML_CART $_{modified}$ on each node

During the construction of the trees, we use a modified version of the GML_CART that fits the randomization requirement of random forests. Thus, Algorithm 3 was developed so that the decision tree nodes use a random subset of the attributes drawn from the original set of attributes without replacement. Following the studies carried out to determine the optimal number of features considered at each split point in the case of classification, we set

this hyperparameter to $\sqrt{d}$ [24, 25].

Finally, in the classification phase, we combine the predictions from the constructed trees and aggregate the predictions of the membership degrees using a label-wise majority vote.

## 4. Experimental study

### 4.1 Evaluation metrics

We evaluated the predictive performance of our models using three evaluation metrics from the GMLC literature, i.e., hamming loss, vertical 0-1 loss, and C-Index.

The hamming loss quantifies the mean deviation of the predicted degrees to the actual ones [20]. It is defined as follows:

$$\text{HAMMIG LOSS} = \frac{\sum_{i=1}^{k} AE(\hat{y}_i, y_i)}{(m-1)k} \quad (6)$$

where $AE$ is the absolute error of the predicted grades:

$$AE: M \times M \rightarrow \mathbb{N},\ AE(\mu_i, \mu_j) = |i-j| \quad (7)$$

The hamming loss can also be defined as the averaged sum of the Manhattan distances between the predicted vectors and the actual ones.

The vertical 0-1 loss represents the percentage of class labels that were assigned the wrong degree. Contrary to the hamming loss, the vertical 0-1 loss only considerers whether a degree is the exact match to the real one and does not take into account the difference between the two. In [19], it is defined as follows:

$$\text{VERTICAL } 0-1 \text{ LOSS} = \frac{1}{k}\sum_{i=1}^{k} I(\hat{y}_i \neq y_i) \quad (8)$$

where $I$ is the indicator function.

This measure can be considered as inversely proportional to the overall accuracy of the model.

The C-Index is a loss that accounts for the pairwise errors in the ranking of the labels. It is defined in [20] as follows:

$$\text{C\_INDEX} = \frac{\sum_{i<j}\sum_{(\lambda,\lambda\prime)\in M_i \times M_j} S([H]_\lambda, [H]_{\lambda\prime})}{\sum_{i<j}|M_i| \times |M_j|} \quad (9)$$

where $M_i = \{\lambda \in L | L_x(\lambda) = \mu_i\}$

$L_x(\lambda)$ is a function that returns the degree of membership of the label for an instance $x$ and

$$S(u,v) = I(u > v) + \frac{1}{2} I(u = v) \quad (10)$$

### 4.2 Datasets

The datasets considered in our experiments include the BeLa-E 100 benchmark datasets and the dataset movies. Table 2 details their properties.

BeLa-E was first introduced in [20] and became a benchmark dataset for GMLC. This dataset was originally obtained from a social psychology study on career development of university graduates [26]. It consists of 1930 instances where each instance represents a graduate student and is described by 50 attributes. Two of these attributes indicate the student's age and gender and the remaining features characterize the degree of importance the student attaches to different aspects of a future job.

In [20], the authors generated 100 variants of BeLa-E by considering random subsets as target labels. 50 of these datasets predict 5 class labels and the other 50 predict 10 labels.

The dataset movies [19] was extracted from a German website that categorizes movies by their degrees of membership or belonging to different genres (action, comedy, suspense...). Rather than assigning a number of genres to a movie, this dataset assigns a level of relevance for each category. These levels range contiguously from '0' to '3'. The number of movies comprised in the data is 1967 movies described by a total of 27002 attributes.

### 4.3 Results

We conducted a comparative study using the aforementioned datasets and evaluation measures. This experimental study included different approaches from the literature and the new developed methods using 10-fold cross-evaluation. We should also mention that the adapted models are developed from scratch using python while the transformation-based models were deployed using the Weka framework [27]. Tables 3-5 summarize the results obtained in this study.

We can notice that in terms of hamming loss and vertical 0-1 loss, GML_RF outperformed the rest of the seven methods used in this experimental study. This means that the adapted random forest proposed has the best predictions of the label grades. On the other hand, we find that the Joined CLR results in better rankings of the labels which are measured by the C-Index. Despite the lower C-Index results for the Joined CLR in comparison to those of the GML_RF, the latter remains a better model considering the significantly higher losses of the joined CLR regarding the rest of the performance measures, especially those of the vertical 0-1 loss where we find more than 26% increase in loss.

Table 2. General characteristics of the graded multi-label datasets: number of variant datasets, number of instances, number of attributes, number of labels, and number of grades

| Datasets | Number of datasets | Instances | Attributes | Labels | Grades |
|---|---|---|---|---|---|
| BeLa-E k=5 | 50 | 1930 | 45 | 5 | 5 |
| BeLa-E k=10 | 50 | 1930 | 40 | 10 | 5 |
| Movies | 1 | 1967 | 27002 | 5 | 4 |

Table 3. Results of the experimental study on 50 BeLaE_k5 datasets in terms of hamming loss, vertical 0-1 loss, and C-Index

| Models | Hamming Loss (%) | Vertical 0-1 Loss (%) | C-Index (%) |
|---|---|---|---|
| **BR** | 25.74 ± 2.61 | 68.95 ± 3.28 | 37.44 ± 5.57 |
| **IBLR-ML [20]** | 27.23 ± 4.51 | 69.39 ± 5.39 | 49.55 ± 8.44 |
| **Full CLR [19]** | 33.97 ± 5.79 | 73.44 ± 7.58 | 20.38 ± 4.13 |
| **Joined CLR [19]** | 17.96 ± 1.31 | 61.82 ± 3.61 | **18.16** ± 3.68 |
| **Horizontal CLR [19]** | 15.77 ± 1.53 | 51.90 ± 3.52 | 23.88 ± 4.11 |
| **GML_DT [21]** | 16.89 ± 1.84 | 52.62 ± 3.90 | 26.43 ± 4.71 |
| **GML_CART** | 16.72 ± 1.76 | 52.40 ± 3.85 | 26.37 ± 4.80 |
| **GML_RF** | **14.76 ±** 1.59 | **47.63 ±** 3.59 | 23.15 **±** 4.50 |

Table 4. Results of the experimental study on 50 BeLaE_k10 datasets in terms of Hamming Loss, Vertical 0-1 Loss, and C-Index

| Models | Hamming Loss (%) | Vertical 0-1 Loss (%) | C-Index (%) |
|---|---|---|---|
| **BR** | 25.97 ± 1.76 | 69.02 ± 2.39 | 36.15 ± 4.05 |
| **IBLR-ML [20]** | 27.27 ± 3.83 | 69.95 ± 4.16 | 50.37 ± 6.98 |
| **Full CLR [19]** | 35.44 ± 3.70 | 75.11 ± 4.47 | 18.57 ± 2.27 |
| **Joined CLR [19]** | 17.92 ± 0.87 | 61.76 ± 0.87 | **17.58** ± 2.14 |
| **Horizontal CLR [19]** | 15.13 ± 0.95 | 50.45 ± 2.15 | 22.78 ± 2.53 |
| **GML_DT [21]** | 17.43 ± 1.10 | 54.06 ± 2.03 | 26.37 ± 3.02 |
| **GML_CART** | 17.32 ± 1.08 | 53.70 ± 2.04 | 26.37 ± 3.10 |
| **GML_RF** | **15.08 ±** 0.96 | **48.63 ±** 1.89 | 23.34 **±** 2.98 |

We can note that GML_RF resulted in an important decrease in the hamming loss and the vertical 0-1 loss on the ensemble of the benchmark datasets. In fact, for the BeLaE k=5 benchmark, we have a decrease in the hamming loss that goes from 56.54% to 6.40% and a diminution in vertical 0-1 loss that ranges from 35.14% to 8.22%.

For the BeLaE k=10 benchmark, we found a drop in hamming loss that varies from 57.44% to 0.33% and a decrease in the vertical loss that ranges from 35.25% to 3.60%. As per the movies dataset, the reduction in hamming loss goes from 2.78% to reach 78.12% and the decrease in vertical 0-1 loss varies from 58.03% to 4.52%.

The registered improvement of GML_RF in accuracy and precision measures rather than ranking can be explained by the aggregation method used in the prediction. The algorithm relies on a label-wise

Table 5. Results of the experiments on the movies dataset in terms of hamming loss, vertical 0-1 loss, and C-Index

| Models | Hamming Loss (%) | Vertical 0-1 Loss (%) | C-Index (%) |
|---|---|---|---|
| **BR** | 25.39 | 53.63 | 36.88 |
| **IBLR-ML [20]** | 32.33 | 67.34 | 33.98 |
| **Full CLR [19]** | 76.51 | 96.50 | 15.43 |
| **Joined CLR [19]** | 25.32 | 67.16 | **14.74** |
| **Horizontal CLR [19]** | 17.73 | 44.70 | 21.40 |
| **GML_DT [21]** | 17.22 | 42.42 | 24.72 |
| **GML_CART** | 17.36 | 42.80 | 25.24 |
| **GML_RF** | **16.74** | **40.50** | 22.22 |

majority vote to predict the degrees of membership instead of considering each prediction vector as a whole. We opted for this approach because it results in higher accuracy, which we consider more important for GMLC.

The GML_CART model proposed in this article slightly improved the results obtained with GML_DT for the BeLa-E benchmark. However, GML_DT regained an edge on the Movies dataset which we explain by the lower feature importance scores on the Movies dataset found in our exploratory analysis of the datasets. Although we should also mention that GML_CART was computationally faster which is explained by the use of logarithms in the entropy in GML_DT.

## 5. Conclusion

In this paper, we presented two algorithms that we adapted to GMLC by modifying the formula of the Gini-Index. This adaptation of the impurity measure allows it to calculate the averaged impurity with respect to all the labels simultaneously and therefore takes into account the dependencies between the labels. To avoid the drawback of overfitting in decision trees, we developed a random forest based on the generated graded multi-label trees. We carried out a comparative study of the proposed models and different approaches from the literature in both the transformation-based methodology and the adaptation-based paradigm. We used three performance measures of GMLC to evaluate the efficiency of the algorithms. The results of the experimental study show the competitive performance of the newly adapted models in general and the overall higher efficiency of GML_RF compared to the rest of the models according to the hamming loss and the vertical 0-1 loss performance measures. In future work, we intend to optimize these methods and further contribute to the evolving research on GMLC.

## Conflicts of interest

The authors declare no conflict of interest.

## Author contributions

Wissal Farsal provided the idea, technique, software, formal analysis, data collection, writing and writing-original draft preparation. Mohammed Ramdani and Samir Anter provided supervision, and editing.

## References

[1] M. Han, H. Wu, Z. Chen, M. Li, and X. Zhang, "A survey of multi-label classification based on supervised and semi-supervised learning", *International Journal of Machine Learning and Cybernetics*, pp. 1-28, 2022.

[2] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods", *Expert Systems with Applications*, Vol. 203, p. 117215, 2022.

[3] N. Endut, W. A. F. W. Hamzah, I. Ismail, M. K. Yusof, Y. A. Baker, and H. Yusoff, "A Systematic Literature Review on Multi-Label Classification based on Machine Learning Algorithms", *TEM Journal*, Vol. 11, No. 2, p. 658, 2022.

[4] Z.H. Zhou and M.L. Zhang, "Multi-label learning", *Encyclopedia of Machine Learning and Data Mining*, pp. 875–881, 2017.

[5] H. E. Rifai, L. A. Qadi, and A. Elnagar, "Arabic text classification: the need for multi-labeling

systems", *Neural Computing and Applications*, Vol. 34, No. 2, pp. 1135-1159, 2022.

[6] H. Lotf and M. Ramdani, "MaroBERTa: Multilabel Classification Language Model for Darija Newspaper", In: *Proc. of Smart Applications and Data Analysis: 4th International Conference*, SADASC 2022, Marrakesh, Morocco, pp. 388-401, 2023.

[7] L. Qing, W. Linhong, and D. Xuehai, "A novel neural network-based method for medical text classification", *Future Internet*, Vol. 11, No. 12, pp. 255, 2019.

[8] W. Zhou, Z. Xia, P. Dou, T. Su, and H. Hu, "Aligning image semantics and label concepts for image multi-label classification", *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 19, No. 2, pp. 1-23, 2023.

[9] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: a multi-label hate speech detection dataset", *Complex & Intelligent Systems*, Vol. 8, No. 6, pp. 4663-4678, 2022.

[10] W. Tang, R. Dai, W. Yan, W. Zhang, Y. Bin, E. Xia, and J. Xia, "Identifying multi-functional bioactive peptide functions using multi-label deep learning", *Briefings in Bioinformatics*, Vol. 23, No. 1, p. bbab414, 2022.

[11] X. Du and J. Hu, "Deep multi-label joint learning for RNA and DNA-binding proteins prediction", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[12] W. Farsal, M. Ramdani, and S. Anter, "A Novel Graded Multi-label Approach to Music Emotion Recognition", In: *Proc. of Smart Applications and Data Analysis: 4th International Conference*, SADASC 2022, Springer, Marrakesh, Morocco, pp. 187-197, 2023.

[13] K. Laghmari, C. Marsala, and M. Ramdani, "An adapted incremental graded multi-label classification model for recommendation systems", *Progress in Artificial Intelligence*, Vol. 7, pp. 15-29, 2018.

[14] K. Laghmari, C. Marsala, and M. Ramdani, "A Distributed Recommender System Based on Graded Multi-label Classification", In: *Proc. of Networked Systems: 5th International Conference*, NETYS 2017, Marrakech, Morocco, pp. 101-108, 2017

[15] G. Lastra, O. Luaces, and A. Bahamonde, "Interval prediction for graded multi-label classification", *Pattern Recognition Letters*, Vol. 49, pp. 171-176, 2014.

[16] F. Charte, A. J. Rivera, and M. J. D. Jesus, "Multilabel classification: problem analysis, metrics and techniques", *Springer International Publishing*, 2016.

[17] M. L. Zhang, Y. K. Li, X. Y. Liu, and X. Geng, "Binary relevance for multi-label learning: an overview", *Frontiers of Computer Science*, Vol. 12, pp. 191-202, 2018.

[18] K. Laghmari, C. Marsala, and M. Ramdani, "Learning Label Dependency and Label Preference Relations in Graded Multi-label Classification", *Computational Intelligence for Pattern Recognition*, pp. 115-164, 2018.

[19] C. Brinker, E. L. Mencía, and J. Fürnkranz, "Graded multilabel classification by pairwise comparisons", In: *Proc. of 2014 IEEE International Conference on Data Mining*, pp. 731-736, 2014.

[20] W. Cheng, E. Hüllermeier, and K. J. Dembczynski, "Graded multilabel classification: The ordinal case", In: *Proc. of the 27th International Conference on Machine Learning (ICML-10)*, pp. 223-230, 2010.

[21] W. Farsal, M. Ramdani, and S. Anter, "GML_DT: A Novel Graded Multi-label Decision Tree Classifier", *International Journal of Advanced Computer Science and Applications*, Vol. 12, No. 12, 2021.

[22] L. Breiman, *Classification and regression trees*, Routledge, 2017.

[23] G. Biau and E. Scornet, "A random forest guided tour", *Test*, Vol. 25, pp. 197-227, 2016.

[24] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 9, No. 3, p. e1301, 2019.

[25] S. Bernard, L. Heutte, and S. Adam, "Influence of hyperparameters on random forest accuracy", In: *Proc. of Multiple Classifier Systems: 8th International Workshop MCS*, Reykjavik, Iceland, pp. 171-180, 2009.

[26] E. A. Brehm and M. Stief, "Die Prognose des Berufserfolgs von Hochschulabsolventinnen und-absolventen: Befunde zur ersten und zweiten Erhebung der Erlanger Längsschnittstudie BELA-E", *Zeitschrift Für Arbeits-und Organisationspsychologie A&O*, Vol. 48, No. 1, pp. 4-16, 2004.

[27] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes, "Meka: a multi-label/multi-target extension to weka", *Journal of Machine Learning Research*, Vol. 17, No. 21, p. 1–5, 2016.