Noname manuscript No.
(will be inserted by the editor)

A New Framework of Multi-objective Evolutionary Algorithms for Feature Selection and Multi-label Classification of Video Data

Gizem Nur Karagoz · Adnan Yazici · Tansel Dokeroglu · Ahmet Cosar

Received: date / Accepted: date

Abstract There are few studies in the literature to address the multi-objective multi-label feature selection for the classification of video data using evolutionary algorithms. Selecting the most appropriate subset of features is a significant problem while maintaining/improving the accuracy of the prediction results. This study proposes a framework of parallel multi-objective Non-dominated Sorting Genetic Algorithms (NSGA-II) for exploring a Pareto set of non-dominated solutions. The subsets of non-dominated features are extracted and validated by multi-label classification techniques, Binary Relevance (BR), Classifier Chains (CC), Pruned Sets (PS), and Random k-Labelset (RAkEL). Base classifiers such as Support Vector Machines (SVM), J48-Decision Tree (J48), and Logistic Regression (LR) are performed in the classification phase of the algorithms. Comprehensive experiments are carried out with local feature descriptors extracted from two multi-label data sets, the well-known MIR-Flickr dataset and a WMS (Wireless Multimedia Sensor) dataset that we have generated from our video recordings. The prediction accuracy levels are improved by 6.36% and 25.7% for the MIR-Flickr and WMS datasets respectively while the number of features is significantly reduced. The results verify that

Gizem Nur Karagoz
Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
E-mail: gizem.karagoz@metu.edu.tr

Adnan Yazici
Department of Computer Science, Nazarbayev University, Nur-Sultan, Kazakhstan
E-mail: adnan.yazici@nu.edu.kz
Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
E-mail: yazici@ceng.metu.edu.tr

Tansel Dokeroglu
Department of Computer Engineering, TED University, Ankara, Turkey
E-mail: tansel.dokeroglu@tedu.edu.tr

Ahmet Cosar
Department of Computer Engineering, University of THK, Ankara, Turkey
E-mail: cosar@thk.edu.tr

the algorithms presented in this new framework outperform the state-of-the-art algorithms.

Keywords Multi-label Classification · Multi-objective Optimization · Evolutionary · Machine learning · Feature Selection

1 Introduction

We live in an era where computer systems produce very large amounts of data that must be processed to extract hidden knowledge. To make the best use of available computing resources, the processing time of big data needs to employ special data processing techniques. Some parts of the data may be contaminated and this can prevent the extraction of useful knowledge. Irrelevant and/or redundant data must be eliminated, preferably even before being transmitted to a big data store, to reduce the data processing load, to increase the classification accuracy and to obtain better data models. Complex data structures need to be designed for filtering out irrelevant data. Efficient data mining and machine learning methods are being used for discovering correlations [1]. The feature selection is one of the most suitable methods to search for the most relevant and the smallest subset of features for the data classification. There are three main methods in literature for performing feature selection: filtering, wrapper, and embedded methods. The filtering method uses computationally inexpensive evaluation functions over all available data features, providing a ranking of features [2]. The wrapper method uses learning algorithms to determine the most relevant subsets to maximize the performance of learning. The evaluation of the wrapper algorithm is computationally very expensive but it determines the most valuable subsets of features [3]. The embedded method combines feature selection techniques with the model construction process (wrapper) so that it can stop the attribute filtering process when sufficient performance is obtained by the classification/learning algorithms [4].

Fig. 1 a. Binary classification, b. Multi-class classification, c. Multi-Label classification

Most of the time, real data includes multiple scopes. An image taken by a camera can contain many features. Tagging such data-rich content with a simple binary label may not always be possible. For this reason, multi-label classification is an important aspect of data classification problems [5][6]. The

data is labeled with one of two classes in binary classification. For the multi-class classification, there are more than two classes and each row of data is tagged only with a single label value. For the multi-label classification, there are more than two classes and each data may have more than one label. In Figure 1, the examples of binary classification, multi-class classification, and multi-label classification of a set of instances are presented.

We handle the problem of multi-label classification of video data problem as a two-dimensional optimization problem. There may be subsets of solutions with a minimum number of features. However, their accuracy may not always be the best. We intend to obtain the smallest feature sets to reduce the execution time of large datasets while improving the classification accuracy of the datasets.

In this study, we propose a framework of multi-objective evolutionary algorithms for the solution of this important multi-objective problem. The well-known NSGA-II algorithm is used in the feature selection phase of the developed algorithms [7]. Later, the multi-label classification machine learning technique validates this set of selected features for their prediction accuracy performance. Binary Relevance (BR), Classifier Chains (CC), Pruned Sets (PS) and Random k-Label Sets (RAkEL) are used in the proposed algorithms since they are the best-performing algorithms used in literature. Support Vector Machines (SVM), J48-Decision Tree (J48) and Logistic Regression (LR) are used to calculate the fitness values. Since the most time-consuming part of these multi-objective evolutionary algorithms is the evaluation of the fitness value of each chromosome, these computations are performed in a parallel computing (multi-threaded) environment to produce speed-up and scalability while examining larger sets of features. We develop twelve algorithms to verify that it is possible to obtain better prediction accuracy with a minimum number of features. The algorithms are validated with two different multi-label image/video datasets. The first one is the well-known MirFlickr dataset [8] with extracted features [9] and the second dataset is a new Wireless Multimedia Sensor(WMS) video dataset produced in our research project (available on our website [1]). The second dataset is manually annotated and the bag-of-visual-words are generated from a local Scale Invariant Feature Transformation (SIFT) descriptor. When the parallel versions of the proposed algorithms are used, the Hamming score values of the individuals in the population are increased and the number of features is decreased significantly.

To the best of our knowledge, we propose/design the first version of the multi-objective multi-label classification problem for the video-datasets in literature. Multi-objective evolutionary algorithms (NSGA-II) have been successfully applied to many feature selection problems. However, the multi-objective parallel feature selection on local descriptors for the image/video datasets has been performed for the first time in this study. Twelve different combinations of the algorithm are developed using multi-threaded programming in our proposed framework. Significant performance improvements are observed and

[1] http://ceng.metu.edu.tr/tr/node/3612

both objectives (higher Hamming score values with the minimum number of features) are achieved concerning the results of the state-of-the-art algorithms.

Section 2 reviews recent studies related to this research. In Section 3, the definition of the problem is given. The proposed framework algorithm for the selection of features is described in detail and the validation algorithms are explained in Section 4. The experimental results of the algorithms are compared and discussed in Section 5. Our concluding remarks and future work are presented in the last section.

2 Related work

This section gives information about the multi-objective evolutionary algorithms that are proposed for feature selection. Guyon et al. present a feature selection method that consists of a heuristic checklist that provides a basic roadmap by asking questions about features and labels [10]. Using the responses, a filter, a wrapper, or an embedded method is decided. Another feature selection research is presented by Jing et al. by using multi-objective optimization algorithms [11]. A multi-label k-nearest algorithm is implemented and tested by four multi-label datasets using Hamming loss parameter. Their method integrates a Genetic Algorithm (GA) with machine learning techniques. They report better results than traditional feature selection algorithms. Zhang et al. implement a new multi-label feature selection method for the classification of data by using a multi-objective Particle Swarm Optimization (PSO) algorithm [12][13]. The introduced algorithm is compared to NSGA-II. Datasets used for validation and comparison have a maximum number of 14 labels and 294 features.

Vaishali et al. propose an evolutionary feature selection algorithm using a multi-objective evolutionary method [14]. For the evaluation phase, four different machine learning algorithms are used Naive Bayes (NB), J48 Decision Tree, MLP, Neural Network and Multi-objective Fuzzy Classification. The NSGA-II and Evolutionary Non-Dominated Radial Slots based algorithm (ENORA) are used in the experiments. During the validation phase, the binary-labeled health dataset with 8 features and 767 instances is used. Both ENORA and NSGA-II algorithms give better results after feature selection but NSGA-II's Hamming score improvement is observed to be better when compared to all other algorithms. Vignolo et al. eliminate irrelevant, noisy and redundant features for the face recognition problem [15]. The implemented evolutionary multi-objective method aims to minimize the cardinality while maximizing the discriminative capacity. The multi-objective (MOGA) and classical GAs are compared and both offer similar accuracy, whereas MOGA achieves this with fewer features.

A study based on the filter method is provided by Labani et al. [16]. In this study, NSGA-II algorithm is used with NB and SVM for feature selection. 5 different binary-class datasets are used. They compare their proposed method with other filter-based methods like Max-Relevance Min-Redundancy and concerning both a minimum number of features and the maximum accu-

racy. The proposed algorithm outperforms commonly used methods. Zhang et al. propose a feature selection approach based on the weighted relevancy [17]. They observe that the correlation between candidate features and class labels have an important role in feature selection. While calculating the relevance between features and class labels; entropy and mutual information are calculated. Deniz et al. propose three feature selection methods for binary classification problems with machine learning techniques [18]. They propose techniques with two phases; feature subset selection and applying machine learning techniques for the prediction accuracy. Saroj & Jyoti work with NSGA-II algorithm for obtaining optimal multi-objective feature selection [19]. One-point crossover, bit-flip mutation, and binary-tournament selection methods are used. The fitness function is determined through an equally weighted sum of objectives which are maximizing information gain, maximizing non-redundancy, and minimizing the feature set. Better results are provided by getting Pareto-optimal solutions instead of a single best solution. Hamdani et al. perform experiments on NSGA-II algorithm for multi-objective feature selection [20]. The 1-NN algorithm is used as a classifier for evaluating the solutions.

Khan et al. [21] propose a multi-objective feature selection algorithm for multi-label data classification. NSGA-II algorithm is used with SVM as a base learner for the fitness values of the algorithm. 2 multi-label datasets with 7 and 174 label values are examined with state-of-art multi-label classifiers, Label Powerset (LP), BR, CC) and Calibrated Label Ranking (CLR). Shijin et al. propose a hybrid method of GA and SVM on feature selection for hyperspectral image classification to get better band combination means finding irrelevant band combinations with a minimal number of bands [22]. Gaspar implements a feature selection algorithm using multi-objective evolutionary methods [23].

Xue et al. propose a PSO for the multi-objective feature selection [24]. Linear Forward Selection (LFS) and Greedy Step-wise Backward Selection (GSBS) methods are used in the proposed algorithm. NSGA-II, Strength Pareto Evolutionary Algorithm-2 (SPEA2) and Pareto Achieved Evolutionary Strategy (PAES) algorithms are compared. The K-nearest neighbor algorithm with 10-fold cross-validation is used in the experiments. Better performance is observed with LFS than GSBS for both the number of features and the prediction accuracy. Zhang et al. propose a similar method that is focused on the performance metrics of multi-objective optimization algorithms [25]. Hyper-volume and two-set-coverage are investigated in this study. The results of the NSGA-II algorithm are reported to be worse than the PSO, whereas NSGA-II has better results than the multi-objective differential evolution feature selection algorithm to the hyper-volume. In a two-set-coverage metric, the NSGA provides better Pareto optimal solutions than the proposed algorithm on the majority of the datasets. Tangherloni et al. investigate the performance of meta-heuristics for real-world optimization problems [26]. The authors study the Parameter Estimation (PE) of biochemical systems, a common computational problem in the field of Systems Biology. They compare the solution quality of their algorithms by considering a set of benchmark functions and a set of biochemical models with an increasing number of di-

mensions. Experimental results verify that some state-of-the-art optimization methods are characterized by considerably poor performances when applied to the PE problem.

Nalluri et al. propose a hybrid architecture, monarch butterfly optimization (MBO), to handle imbalanced binary disease datasets that arrive upon the efficient combination of SVM classifiers sensitive parameter values of evolutionary algorithms [27]. MBO enumerates three objectives, prediction accuracy, sensitivity, specificity. A uni-modular matrix and limit points based non-dominated solutions selection for local and global search and to generate an efficient initial population respectively are introduced. The performance of the architecture is verified on 18 disease datasets having binary class labels and significant improvements are obtained.

Rundo et al. propose a novel image enhancement method (MedGA) based on Genetic Algorithms to improve the appearance and the visual quality of images characterized by a bimodal gray-level intensity histogram, by strengthening their two underlying sub-distributions [28][29]. MedGA improves the results achieved by downstream image processing techniques. The performance of MedGA quantitatively outperforms the other state-of-the-art tools in terms of signal and perceived image quality while preserving the input mean brightness.

In this study, twelve multi-objective algorithms and their parallel versions are developed and verified. In this sense, our study is unique when compared with other related studies.

3 Problem Definition

In this section, we give information about the multi-label data classification and the multi-objective optimization problems.

3.1 Multi-label classification

The classification problem can be depicted as: let $\mathcal{D}$, $\mathcal{Y}$, and $\mathcal{H}$ be the domain of possible training instances, the class labels, and the set of classifiers respectively. Then, each instance $d \in D$ is assigned to a value $y \in Y$ to find a classifier $h \in H$ that gives maximum possible probability of satisfying $h(d) = y$ for each test case (d, y) [30].

The classification process can be considered in three groups; binary, multi-class, and multi-label classification. In binary classification, the list of class labels Y contains only two values and the classification operation determines whether a test instance belongs to a class or not. In the multi-class classification, the list of Y can have more than two classes for each instance. The membership of these classes is mutually exclusive and there is a unique class label for each instance. For multi-label classification, the list of labels can have multiple values, as in the multi-class classification but each instance can be

assigned to multiple class labels. The classification steps are similar to the binary/multi-class classification processes. The model is first trained with or without labels and then the trained model is validated with distinct data to measure the success of the model. Multi-label classification management requires much effort than other methods. Therefore, some modifications need to be performed. Three main approaches used for this purpose are data transformation, method adaptation and ensemble-based classifiers [31].

Data transformation: can be achieved in two ways. The first one is a transformation into binary classification and the second one is a multi-class classification. For the binary classification transformation, the data is split into single labeled data for all labels and then the results of all classifications are ensembled. For the multi-class classification transformation, label combinations are generated with selected methods. The multi-class classification is applied as if all generated sub-label sets are represented as a single class. Unlike the transformation of binary-classification, the labels are not independent of each other. Their relationships are considered in this method. The most widely used algorithms for this approach are BR, CC, LP, and PS.

Method adaptation: The multi-labeled data remains the same as original but traditional classification algorithms are adapted to handle multiple outputs. There is not a single general solution. Two main concerns are taken into account. The first one is deciding an error function and the second one is a modification for the adaptation of the error function [25].

Ensemble-based classifier: aims to remove barriers or manage weaknesses in methods based on processing or adaption. For example, when applying the PS algorithm, many levels are created and the computation becomes longer. However, the ensemble-based algorithm for PS uses a voting system to avoid a large number of levels. With this modification, the building time is reduced and the success of the algorithm is improved. A detailed explanation of this process is given in [32]. Another popular example is RAkEL that is an ensembled algorithm for LP and BR transformation based algorithms. RAkEL overcomes obstacles produced by LP with two additional parameters, namely the number of classifiers for the training phase and the length of the labelsets. As a result, the performance becomes better than those of BR and LP [33].

3.2 Multi-objective problems

Multi-objective optimization problems aim to obtain two or more objectives related to the solution of the problem. Since the objectives must be related to each other, all objectives are considered simultaneously during the solution. Multi-objective optimization algorithms can be formulated mathematically as

in Equation 1 where f and x represent the objective function and the generated decision space solutions related to M number of objectives respectively [34].

$$f(x) = (f_1(x), f_2(x), ..., f_m(x)) \quad such\ that \quad m = 1, .., M \tag{1}$$

Because the objectives conflict, getting the best results for all objectives is neither meaningful nor possible most of the time [35]. Therefore, Pareto optimality method is used to evaluate multi-objective algorithms. A set of solutions is provided instead of a single best one. Edgeworth proposes the Pareto-optimality paradigm [36]. Concerning this definition of Pareto-optimality, the decision vector is not dominated by any other solution [37].

4 Proposed multi-objective evolutionary algorithm

In this section, we explain the proposed multi-objective evolutionary feature selection algorithms, NSGA-II (the non-dominated sorting GA), the multi-label classification algorithms (BR, CC, PS, RAkEL) and the machine learning techniques (SVM, LR, J48). The algorithms proposed are wrapper type feature selection algorithms.

4.1 Non-dominated Sorting Genetic Algorithm (NSGA-II)

In the first phase of the algorithm, an initial population is randomly generated and the fitness values of individuals (chromosomes) in the population are calculated. The distance of each chromosome's objective values is evaluated with the Euclidean Distance measure called *Crowding Distance*. Since all solutions in the population have two parameters as front and crowding distance, a non-dominated sorting operation is performed by considering the fronts. Individuals in smaller fronts are assumed to have higher priority. For individuals on the same front, their crowding distances are compared and an individual with a larger value is decided to be the winner [34]. The binary tournament selection method is applied to produce the next generations. Four chromosomes are chosen at random. Two parents are selected according to their fitness values. After the crossover and bit-flip mutation operations, two new children are generated. Only the best half of the individuals is used to create the next generation. Individuals with worse fitness values are eliminated. The algorithm terminates when the maximum number of generations is executed [7][27]. In Figure 2, the chromosome structure of an instance with eight features is presented. Genes with value one represent the selected feature indexes of each data instance, whereas the genes with value zero represent features that are not selected for the validation. The flowchart of the proposed algorithm is given in Figure 3. The pseudocode of the proposed algorithm is provided in Algorithm 1.

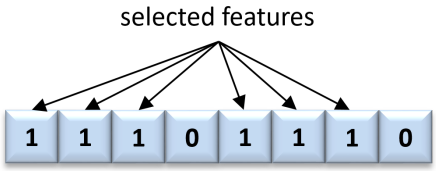

Fig. 2 Chromosome structure of the proposed multi-objective evolutionary algorithms for multi-label image/video classification problem.

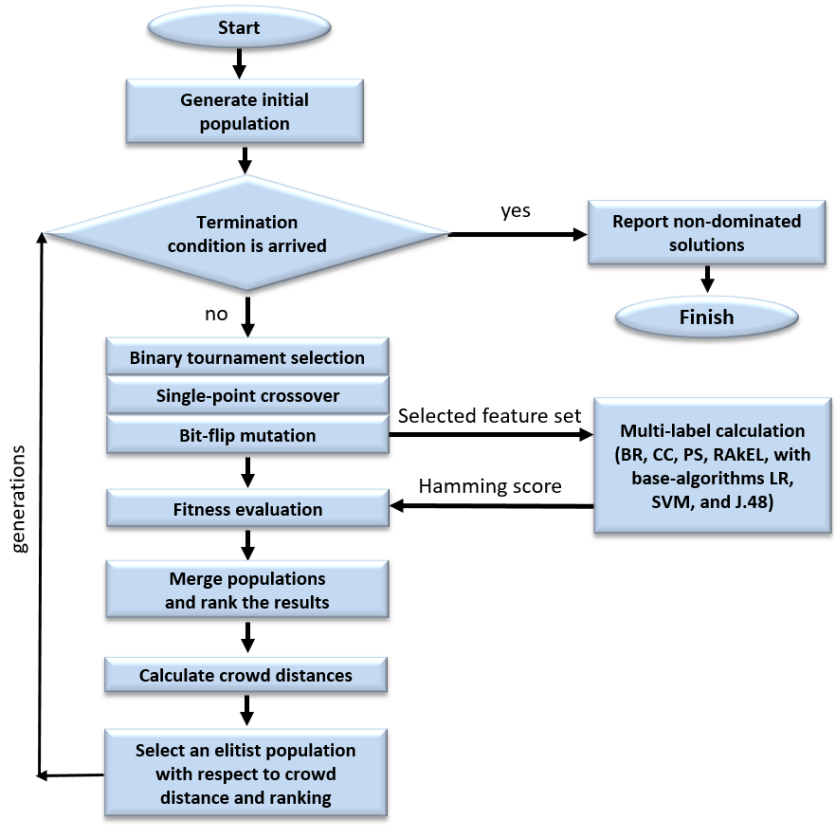

Fig. 3 The flowchart of the multi-objective evolutionary algorithms.

4.2 Multi-label classification algorithms

To handle the multi-labeled datasets, multi-label classification approaches given below are implemented.

Binary Relevance (BR) algorithm: is the multi-tag classification algorithm that is based on the most used transformation to manage multiple targets. The BR divides the data into multiple binary-classification problems and then

Algorithm 1: The pseudocode of the multi-objective evolutionary algorithm for the multi-label data classification problem.

```
Input : popSize, genSize, crossoverRate, mutationRate;
Output : Non_Dominated_Solutions;
P ← randomly generate initial population () ;
S ← {} // the set of already examined individuals/chromosomes ;
for i ← 1 to genSize do
    foreach u in P do
        if u does not exist in S then
            u.objective_1 ← # of selected features;
            // method = {BR,CC,PS,RAkEL with J48,SVM,LR}
            u.objective_2 ← Find_Accuracy(u, method);
            S ← S ∪ {u};
        else
            u.objective_1 ← S[u].objective_1;
            u.objective_2 ← S[u].objective_2;
    P ← Non_Dominated_Sort(P)
    for i ← 1 to |P|/2 do
        p_1 ← Binary_Tournament();
        p_2 ← Binary_Tournament();
        c_1, c_2 ← Half_Uniform_Crossover(p_1, p_2, crossoverRate);
          c_1 ← Bit_Flip_Mutation(c_1, mutationRate);
          c_2 ← Bit_Flip_Mutation(c_2, mutationRate); P ∪ c_1 ∪ c_2;
    P ← Non_Dominated_Sort(P);
return Non_Dominated_Solutions(P);
```

fuses them for the final prediction (see Figure 4). Binarization techniques are used and for each label, a new binary-labeled dataset is created and trained. Test and validation operations are applied to each model for each label in the datasets. Since each target is managed individually, the algorithm does not consider the correlation between labels [31].

Classifier Chains (CC): is similar to the BR algorithm. Multiple binary datasets are generated as shown in Figure 5. When a new binary classification algorithm is executed for every label, the previous labels act as inputs of the next classification process. Unlike BR, the correlation between the labels is not taken into account in CC. Once the data transformation is applied, the classification is managed with binary classifiers for all the generated datasets [38].

Pruned Sets (PS) method: transforms multi-label data into multi-class data. This method is based on the most correlated labels and co-occurrences in multi-labeled datasets. For each combination of label-sets, a new class is registered for a multi-class classification. To handle a large number of classes created, the label set graph is generated based on the condition of the co-occurrences. Thus, the multi-label classification problem can be solved as a

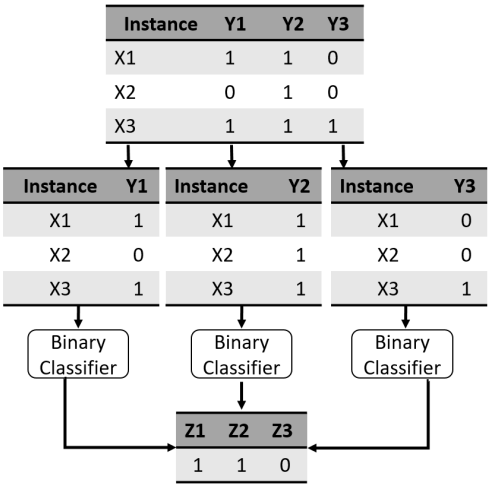

Fig. 4 The stages of the BR algorithm.

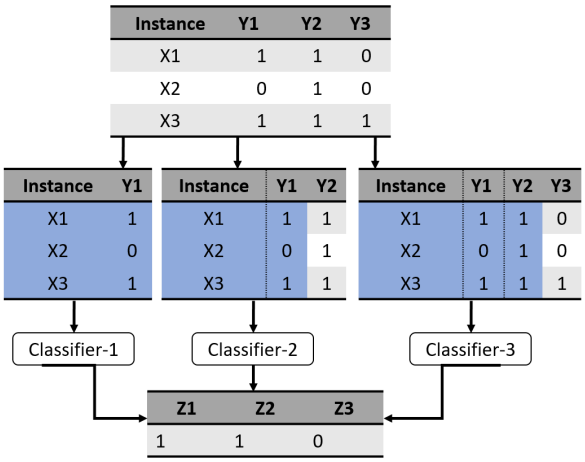

Fig. 5 Data transformation operation of the CC algorithm.

single label classification problem. A new dataset is created as an empty set and all labels with their co-occurrences form a graph with the stopping condition corresponding to the number of label occurrences. If the condition is not fulfilled, these label sets are split and recombined for more common combinations [32].

Random k-Labelsets (RAkEL) method: is an ensemble-based multi-label classification algorithm. Basically, RAkEL creates sub-label sets that consist of k labels. Each sub labelset is selected randomly. RAkEL ensembles these k-

labeled subsets with a label power-set algorithm. Label powerset algorithm is a transformation based multi-label classification technique that takes all labelset as a label of the multi-class classification task. The correlation between labels is taken into account. However, as data size grows, the number of classes and datasets created becomes prohibitively large. Theoretically, it is limited by the number of instances (since $2^k >> n$).

4.3 Performance evaluation

The evaluation metrics used for multi-class or binary classification cannot be used directly for multi-label classification. The accuracy of the labels must be taken into account in the label set. In this way, Hamming loss is the sample-based metric that is used primarily. The loss measure is calculated for each instance and the average value is found. The symmetric difference (Δ) is calculated between the prediction and the actual label sets for all labels per instance (Equation 2). Then, it is normalized according to the number of instances and the number of labels [31].

$$Hamming\ Loss = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n} \mid Y_i \Delta Z_i \mid \tag{2}$$

4.4 Machine learning algorithms

Some basic classifiers are needed to apply multi-label classification algorithms. The classification performance on machine learning algorithms refers to the success of the dimension reduction phase performed by the evolutionary GA.

Support Vector Machines (SVM): is a classification algorithm that creates hyper-planes for the separation of class instances. The hyper-planes are calculated to decide the best hyper-plane maximizing distances from all instances. Hyper-planes are selected according to this condition iteratively.

If data is linearly separable, this method performs well. Otherwise, the kernel trick is used to transform the data into higher dimensionality. It helps to separate the data with hyper-planes [39]. In this study, the Sequential Minimal Optimization (SMO) algorithm is used for this purpose. This algorithm is created by Platt to avoid working on very large data and time-consuming processes [40]. SMO divides the problem into smaller sub-problems. The required memory for SMO is linear and larger training operations can be performed easily. The optimization stage is performed with Lagrange multipliers.

Logistic Regression (LR): is a well-known classification algorithm for both statistics and machine learning. The LR classification algorithm is based on calculating posterior probabilities of occurrence for the attributes in a training set. The Sigmoid function (Equation 3) is used for calculating the probabilities

in our problem domain. X is the input set and θ is the coefficient value for all features. With these values, y gives the probability of the occurrence of an event. Since the label value prediction must be in a binary form, these probability results that are calculated with the Sigmoid function are transformed into binary forms. If the probability of occurrence is less than 0.5, it is assumed to be zero, otherwise one.

$$P = (y = 1 \mid X, \theta) = \frac{1}{1 + e^{-}(\theta * X)} \tag{3}$$

Continuous real values are converted into binaries. If the probability of occurrence is less than 0.5, we take as zero otherwise, we take one. Therefore, LR is a convenient technique for binary classification process [41].

Decision Tree (J48): is a uni-variate algorithm that generates hyper-planes to create partitions in classes. Thus, the branching of a tree depends on a unique attribute. J48 is based on C4.5 decision tree algorithm which is an extension of ID3 [42]. The ID3 algorithm addresses classification problems by creating simple and small decision trees. The potential for all attributes and gains are calculated.

The entropy of all individuals is calculated with Equation 4 and all possible combinations of conditional entropy are computed with Equation 5. Finally, the gains are calculated with Equation 6 to search for anomalies in the data. Once the gain calculations are complete, the best attribute is selected based on the maximal gain for branching. This operation is repeated until all attributes are in the tree or stop condition is fulfilled. When the tree is formed, it is pruned to obtain a more generalized tree and increase the performance of the classification [43].

$$Entropy(y) = \sum_{j=1}^{n} \frac{(\mid y_i \mid)}{(\mid y \mid)} \log \frac{\mid y_i \mid}{\mid y \mid} \tag{4}$$

$$Entropy(j \mid y) = \frac{\mid y_i \mid}{\mid y \mid} \log \frac{\mid y_j \mid}{\mid y \mid} \tag{5}$$

$$Gain(y, j) = Entropy(y - Entropy(j \mid y)) \tag{6}$$

4.5 Parallel versions of the multi-objective evolutionary algorithms

Parallel multi-objective evolutionary algorithms are efficient tools for the optimization of NP-Hard problems [44][45]. The performance of the optimization can be considerably improved by using fine-grained parallel processing of chromosomes with intelligent operators (mutation and crossover). The fitness calculation of the chromosomes in this study requires a lot of time because of the long execution time of applied machine learning techniques. This process prevents the efficient exploration of the subset of features of selected elements.

Therefore, we propose a Parallel-NSGA-II algorithm that uses multi-threaded paradigm for speeding-up the solution of the problem [46]. The proposed algorithm keeps a population in the memory of the master thread and sends chromosomes to be calculated to the slave threads. Since the calculation of the machine learning accuracy with a selected number of features is fine-grained, it is observed that this parallelization technique of the conventional NSGA-II algorithm provides an almost linear speed-up. It is possible to calculate larger numbers of fitness values and obtain better results than the standard (serial) version of the NSGA-II algorithm. The scalability of the algorithms is one of our main concerns to be able to increase the number of cores in the computation environment. The versions we develop in this framework are proved to be scalable and almost linear speed-up in their executions. Rundo et al. make use of the Message Passing Interface (MPI) specifications for Python Master-Slave paradigm employing mpi4py to leverage High-Performance Computing (HPC) resources in medical image analysis provided by MedGA [28][29]. The results of our experiments are reported in the experimental evaluation section.

5 Performance Evaluation of Experimental Results

In our experiments, two multi-label video/image datasets are used to verify the proposed algorithms. The first dataset is the most widely used and publicly available image dataset MIR-Flickr [8]. This dataset consists of 25,000 images. Important features of the dataset are extracted in a study by Costa et al. [9]. This feature set that is extracted with the Segmentation based Fractal Texture Analysis (SFTA) algorithm is used in our experiments. This extraction creates binary images with binary stack decomposition. Extracted features are transformed into vectors as feature sets [47]. There are 42 features in MIR-Flickr dataset and there are at most 23 labels for each image (Car, Bird, Lake, Night, Water, Sky, People, Baby, Clouds, Tree, Portrait, Dog, Animals, Female, Transport, Flower, Indoor, Male, Food, River, Structures, Sea, Sunset).

The second dataset is obtained from our Wireless Multimedia Sensors (WMS) video recordings. The recorded files are split into five-second frames and a manual annotation process is applied to identify objects in the frames. The objects are grouped using three labels (person, group of people and vehicle). After the annotation process is completed, the features are extracted with SIFT method based on key-point localization of objects [48]. The implementation is provided by using openCV framework and the Python programming language [49]. Once the SIFT features are specified, the codebook is constructed to obtain a dictionary of visual words. During the construction of the codebook, the k-means clustering algorithm is applied to determine the centroids. Then, the L1 normalization is applied to obtain the final version in the form of 100 bags of visual words for each frame. The data is extracted from 3-minute videos. 1000 video frames are used in the experiments. In the dataset of WMS, for every frame, 100 features and 3 labels are available as

Fig. 6 Person, group of people, and vehicle frames from the Wireless Multimedia Sensors (WMS) video recordings.

person, group of people and vehicle. Some examples of these frames are shown in Figure 6.

The experiments are performed on a computer with 8 core 64-bit CPU (i7-3632QM, 2.20GHz). The algorithms are developed with Java and the MOEA framework [46]. Multi-label machine learning algorithms providing fitness values are implemented with MEKA, a multi-label extension of WEKA machine learning toolkit [50].

The results reported here are the average of five executions with five-fold cross-validation. This method is used to minimize the impact of random factors. The dataset is divided into five equal-size partitions and four of them are used for training. The remaining partition is used for testing. The average of these five executions is the final accuracy value of the results. This is the most common way in the literature to evaluate the predictive accuracy of machine learning algorithms.

5.1 Setting the size of the population and the number of generations

The main parameters that affect the performance of the NSGA-II are the number of generations and the size of the population. Tuning these parameters improves the efficiency of the NSGA-II significantly. To obtain the best settings, the size of the population is increased from 10 to 100, while the number of generations is being increased to 10, 20, 30, 40, 50, to 70 iterations respectively. Some of the results are presented in Figures with different number of generations and populations 7, 8, 9 and 10.

At the beginning of the experiments, the population size is set to 10 and optimized through 10 generations. The Hamming score reaches 0.885 with three features. The solutions do not construct a good Pareto-curve with this set of parameters. In advanced stages of experiments, a fewer number of features are achieved with higher Hamming scores. Non-dominated results are well located on Pareto-curves. During the experiments, the increase in the number of instances in each population positively affects the results. Better results

Table 1 Parameter settings for the parallel NSGA-II algorithms (N is the number of individuals in the population).

Parameter	**Value**
population size	50
# generations (termination condition)	20
crossover rate	1.0
distribution index for crossover	15.0
mutation rate	$1/N$
distribution index for mutation	20.0

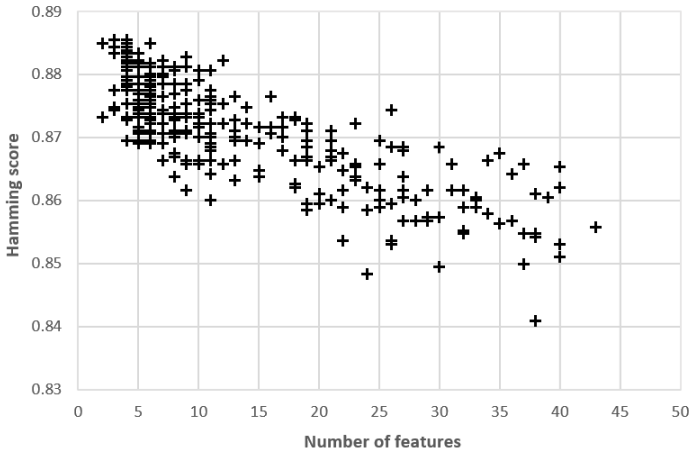

Fig. 7 The performance of the multi-objective BR-J48 algorithm with 10 individuals and 10 generations.

can be achieved as the number of generations increases. The best Hamming score of minimum features - optimal solutions is very similar to each other with 50 population - 70 generations and 100 population - 30 generations. Deciding the number of generations as 20 and the number of instances for each population as 50 is suitable for obtaining the best non-dominant solutions. The best parameters used in the experiments for the parallel NSGA-II are presented in Table 1.

5.2 Experimental results of serial and parallel NSGA-II algorithms

For the performance comparison of serial and parallel versions of the algorithms, four multi-label classification algorithms with base classifier J48 decision trees are tested with the MIR-Flickr dataset. The results are observed to construct better Pareto-curves (see Figures 11, 12, 13 and 14 for more details).

The serial PS algorithm has Hamming score values 0.812 and 0.823 respectively with one and 26 features. The parallel version of this algorithm has 0.840 Hamming score with one feature and the Hamming score is maximized

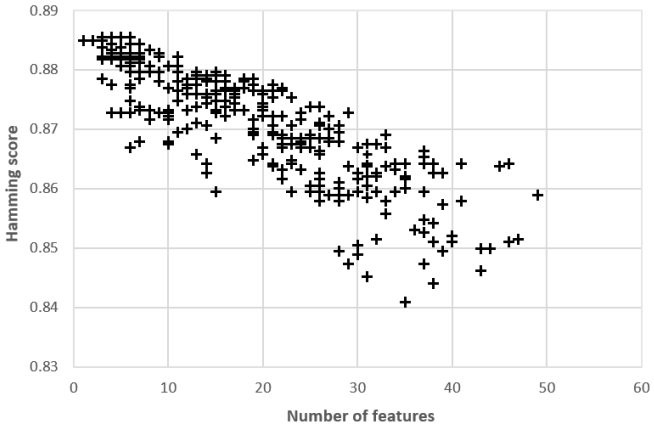

Fig. 8 The performance of the multi-objective BR-J48 algorithm with 30 individuals and 10 generations.

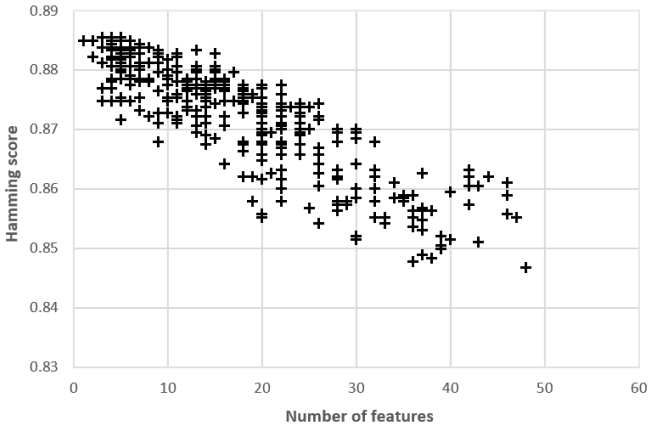

Fig. 9 The performance of the multi-objective BR-J48 algorithm with 10 individuals and 70 generations.

to 0.861 with 7 features. In both objectives, better results are obtained with the parallel versions of the algorithms.

The serial implementation of the CC algorithm has 0.866 Hamming score with one feature. With the same number of features, 0.883 Hamming score is obtained in the parallel version of the algorithm.

The serial implementation of BR has 0.866 Hamming score with one feature. The best non-dominated results (Pareto-optimal results) of the algorithm and its parallel version have 0.887 Hamming score with the same number of features. The maximum Hamming score is 0.889 with 6 features in parallel version and 0.867 with seven features in the serial version.

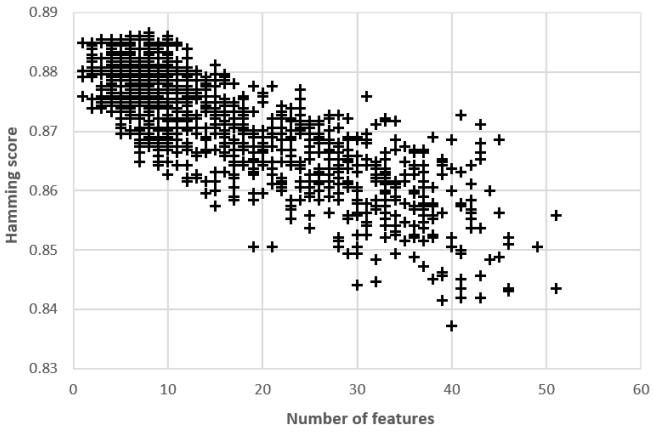

Fig. 10 The performance of the multi-objective BR-J48 algorithm with 50 individuals and 50 generations.

Table 2 The average execution time of serial and parallel algorithms

Algorithm	Base Classifier	serial (min)	parallel (min)	Improvement (%)
BR	J48	371.23	23.28	93.73
CC	J48	389.54	20.31	94.79
PS	J48	1020.6	36.27	96.45
RAkEL	J48	371.36	114.55	69.15

When the resulting graphs are analyzed, all the algorithms are observed to converge faster with their parallel versions. The convergence of the serial algorithm requires more execution time and cannot achieve results of parallel implementations in terms of Hamming scores. The parallel multi-objective evolutionary algorithms save a great amount of execution time. The serial version of the BR algorithm runs for 6 hours and 11 minutes, whereas the parallel version terminates in 23 minutes and 28 seconds. The serial CC algorithm consumes 6 hours and 30 minutes, whereas its parallel version runs for 20 minutes and 31 seconds. As a result, the parallel versions of the algorithms have a considerable advantage over their serial versions in terms of computation time and solution quality (see Table 2). The improvement in the Table presents the percentage of reduction in execution time (See Equation-7). S and P represent the execution time of the serial and parallel implementation respectively.

$$Improvement\ Percentage = (\frac{S-P}{S}) \times 100 \tag{7}$$

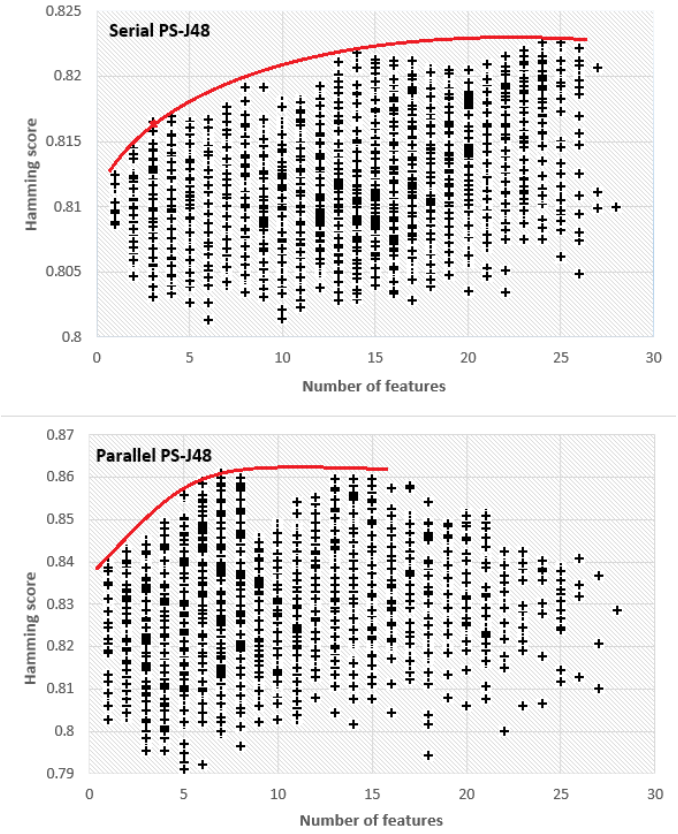

Fig. 11 Comparison of serial and parallel versions of PS-J48 algorithms.

5.3 The evaluation of results with parallel multi-label image/video classification algorithms

Tables 3, 4, 5, and 6 give the performance of BR, CC, PS and J48 multi-label classification algorithms on MIR-Flickr dataset respectively. For each algorithm, 5 Pareto-optimal solutions are reported. After the feature selection process is applied, the Hamming score is improved and the number of features is decreased. Thus, the proposed algorithms are effective in finding relevant features among the set of extracted SFTA features. The best improvement is observed with CC and J48. The Hamming score is improved from 0.8338 to 0.8865. The number of features is decreased from 42 to 9 features. For both objectives, the J48 algorithm outperforms the SVM and LR algorithms. In the original feature set, there are 42 features and it is decreased down to 5 or 9 features in the average. The best Hamming scores are recorded as 0.887 with the CC algorithm. This value was 0.8511 before the feature selection is applied. The RAkEL algorithm gives the worst results concerning the number

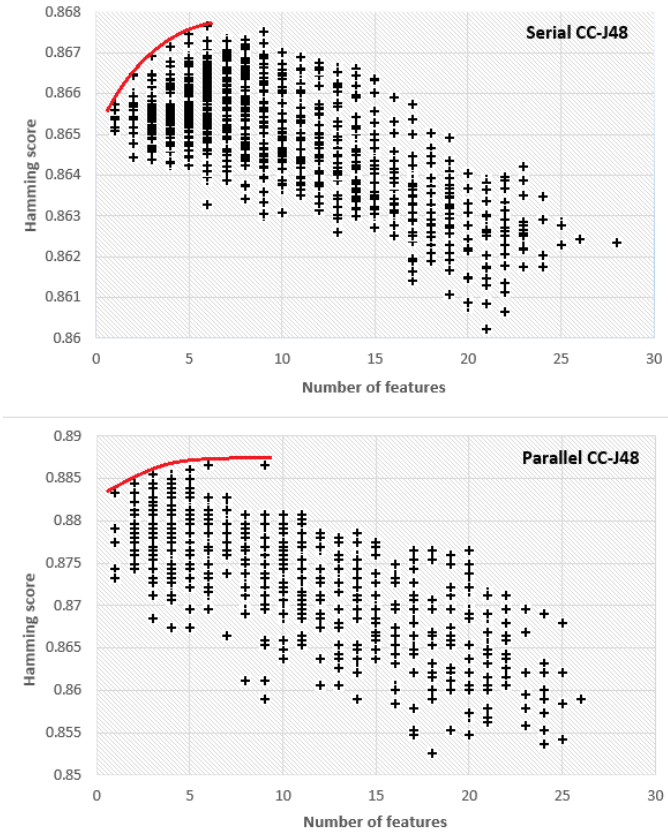

Fig. 12 Comparison of serial and parallel versions of CC-J48 algorithms.

of features compared to other algorithms. Since it finds Pareto-optimal solution in the early stages of the evaluation, the number of non-dominated results is lower than the others. For this dataset, the RAkEL algorithm does not perform well. Our Hamming score results range from 0.80 to 0.84 (similar to other multi-label classification approaches) [51].

Tables 7, 8, 9 and 10 present the results of WMS multi-label video dataset for proposed BR, CC, PS and RAkEL algorithms respectively. The improvement in the Hamming scores is evident after the application of the feature selection process. For the PS-SVM combination, the Hamming score is increased from 0.7038 to 0.8447 and the number of features is reduced from 100 to 41. The best improvement is recorded with the combinations of J48 algorithm. With the BR-J48 algorithm, the Hamming score is improved from 0.6447 to 0.78481 while the number of features is reduced to 14. Similar improvements are obtained in the algorithms PS-J48, CC-J48, and RAkEL-J48. The BR-LR shows a marked improvement with respect to the Hamming score when accuracy results increase from 0.6941 to 0.8456. However, the number

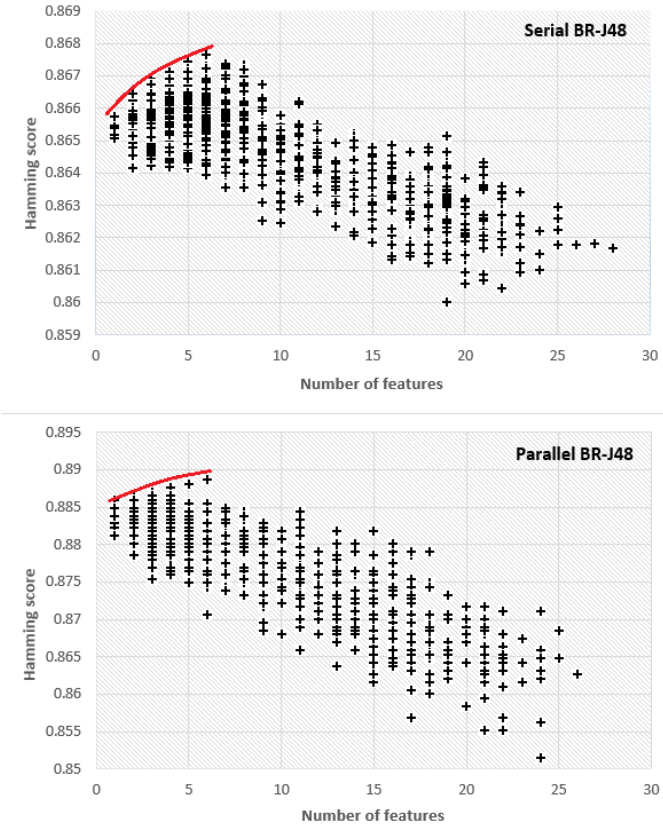

Fig. 13 Comparison of serial and parallel versions of BR-J48 algorithms.

Table 3 Pareto-optimal results of BR algorithm on MirFlicker dataset

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
BR	42	LR	0.86115	0.86939	12
				0.86802	8
				0.86749	6
				0.86668	4
				0.86594	1
		SVM	0.86573	0.87964	25
				0.88335	15
				0.88441	8
				0.88494	2
				0.88494	1
		J48	0.86182	0.88865	6
				0.88812	5
				0.88706	4
				0.88653	2
				0.88602	1

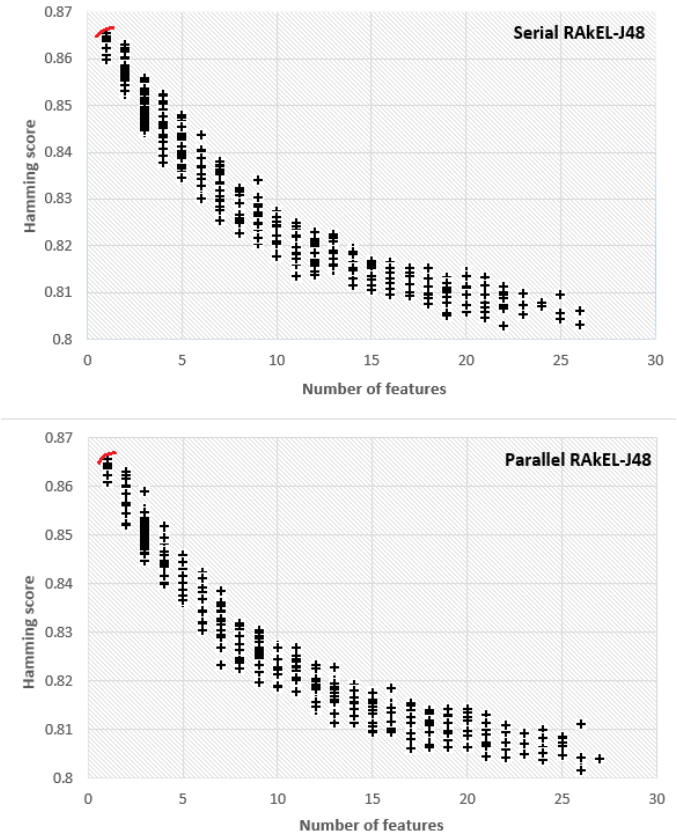

Fig. 14 Comparison of serial and parallel versions of RAkEL-J48 algorithms.

Table 4 Pareto-optimal results of CC algorithm on MirFlicker dataset

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
CC	42	LR	0.84313	0.86532	6
				0.88601	5
				0.88547	3
				0.88441	2
				0.88345	1
		SVM	0.85119	**0.88706**	7
				0.88601	5
				0.88546	3
				0.88442	2
				0.88335	1
		J48	0.83834	0.88653	9
				0.88653	6
				0.88601	5
				0.88494	4
				0.88335	1

Table 5 Pareto-optimal results of PS algorithm on MirFlicker dataset

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
PS	42	LR	0.80201	0.82148	15
				0.82010	8
				0.81933	6
				0.81929	5
				0.81546	1
		SVM	0.84291	**0.86647**	11
				0.86002	9
				0.85790	6
				0.84889	3
				0.83987	1
		J48	0.80799	0.86426	15
				0.86320	9
				0.86108	7
				0.85047	4
				0.83987	1

Table 6 Pareto-optimal results of RAkEL algorithm on MirFlicker dataset

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
RAkEL	42	LR	0.84492	0.80257	28
				0.80419	21
				0.81215	12
				0.83580	5
				0.86439	1
		SVM	0.86541	0.80450	20
				0.81197	13
				0.81894	9
				0.84313	3
				0.86265	1
		J48	0.80232	0.80154	26
				0.80626	17
				0.82253	8
				0.83759	5
				0.86485	1

of features is not decreased as in J48 algorithm. The RAkEL-SVM has the worst performance with respect to both objectives. While the Hamming score is being increased from 0.7038 to 0.7650, the number of features is decreased from 100 to 47. This is the highest number of features recorded after applying feature selection.

To emphasize the importance of feature selection, all combinations of algorithms are executed with the original set of features and three features on the WMS dataset. The execution times of the algorithms are recorded in Table 11. The improvement percentage is calculated with respect to the Equation 7. The percentage of improvement is observed to be remarkably good. The smallest percentage of improvement is 58.41%. But for most algorithms, the improvement is about 95.0%. An improvement of 99.58% is recorded with 3 features.

Table 7 Pareto-optimal results of BR algorithm applied on WMS video dataset.

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
BR	100	LR	0.69409	**0.84557**	32
				0.83587	26
				0.81814	16
				0.76161	5
				0.71856	1
		SVM	0.71645	0.84937	42
				0.84051	34
				0.82278	18
				0.79283	10
				0.74093	3
		J48	0.64472	0.78481	14
				0.78017	11
				0.76835	5
				0.78017	3
				0.74093	1

Table 8 Pareto-optimal results of CC algorithm applied on WMS video dataset.

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
CC	100	LR	0.68776	0.77764	11
				0.77553	8
				0.76624	6
				0.75316	4
				0.72236	1
		SVM	0.71814	0.77595	13
				0.76583	8
				0.75864	5
				0.75063	3
				0.74177	1
		J48	0.64556	**0.78143**	14
				0.77848	10
				0.75751	7
				0.75738	4
				0.74177	1

5.4 Comparison with state-of-the-art algorithms

Our proposed algorithms are compared with state-of-the-art feature selection algorithms, Principal Component Analysis (PCA), Information Gain (IG), and Correlation Based Feature Selection (CBFS). PCA is a linear dimensionality reduction technique that uses linear mapping via covariance or correlation relationship between features. Though variance of the low dimensional data is maximized and by using eigenvectors, most related features arise. This algorithm is based on a study by Pearson [52]. This supervised dimensionality reduction technique is revised in a book by Jolliffe [53]. The other implemented algorithm is IG. This method is used for splitting decision trees but also it is a popular feature selection technique. The difference between the entropy of dataset D and the weighted sum of selected subset entropies are calculated as

Table 9 Pareto-optimal results of PS algorithm applied on WMS video dataset. (FS stands for feature selection.)

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
PS	100	LR	0.66245	0.76709	45
				0.75992	26
				0.74768	13
				0.72152	5
				0.70844	1
		SVM	0.70379	**0.84473**	41
				0.83292	29
				0.81351	16
				0.77722	9
				0.70844	1
		J48	0.62067	0.77004	15
				0.75569	11
				0.74434	8
				0.72532	6
				0.71898	1

Table 10 Pareto-optimal results of RAkEL algorithm applied on WMS video dataset. (FS stands for feature selection.)

Multi-Label Classification Algorithm	# of Features Before FS	Base Classifier	Hamming Score Before FS	Hamming Score after FS	# of Features after FS
RAkEL	100	LR	0.66245	**0.77131**	21
				0.75949	16
				0.75527	10
				0.73038	7
				0.69958	1
		SVM	0.70379	0.76498	47
				0.76287	18
				0.75105	12
				0.74641	8
				0.73881	1
		J48	0.62067	0.76708	38
				0.75949	27
				0.75232	10
				0.73038	5
				0.70465	1

the information gain and the highest is selected as the strongest feature. For this purpose, searching is performed via ranking all attributes. While applying IG on multi-label data, multi-label classification techniques are used. Binary relevance based IG results are evaluated on other multi-label classification algorithms. The last state-of-the-art feature selection algorithm we use is CBFS. It is a filter-based feature selection algorithm and ranks features by a heuristic evaluation function given in Equation 8. The average class-feature correlation is represented as r_{cf}. r_{ff} represents the average feature-feature correlation where k represents the number of features. The subsets are evaluated considering feature-feature and feature-class correlations of all features, termination

Table 11 The improvements in the execution times (with 3-features and 100-features).

Multi-Label Classification Algorithm	Base Classifier	Exec. Time (sec) with all features	Exec. Time (sec) with 3 features	Decrease in Percentage (%)
BR	LR	0.358	0.025	93.02
	SVM	0.194	0.032	83.51
	J48	0.195	0.006	96.92
CC	LR	0.374	0.023	93.85
	SVM	0.188	0.034	81.91
	J48	0.193	0.007	96.37
PS	LR	1.877	0.038	97.98
	SVM	0.188	0.034	67.21
	J48	0.138	0.014	89.86
RAkEL	LR	17.44	0.719	95.88
	SVM	1.878	0.781	58.41
	J48	1.363	0.097	92.88

is performed by the 'best-fit' search method. If five consecutive subsets are not improved over the current best subset then searching is terminated.

$$\mu_s = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \tag{8}$$

State-of-the-art algorithms are evaluated on both datasets. Tables 12 and 13 present the results obtained with MIR-Flickr and WMS datasets respectively. For both dataset BR, CC, PS, and RAkEL multi-label classification algorithms are applied with base classifiers J48 decision tree, SVM and LR on dimensionally reduced subsets. The results show that our proposed algorithm performs better than the-state-of-the-art algorithms in terms of the number of features and Hamming-score values.

With MIR-Flickr dataset, the BR-J48 algorithm has achieved 0.88865 Hamming-score value with six features. With the same algorithm combination, CBFS reports 0.86335 Hamming-score value with 17 features, IG reports 0.86265 with the same number of features and PCA has better results than both CBFS and IG. The results of the CC-SVM and other feature selection algorithms are reported in Figure 15. All state-of-the-art feature selection algorithm results are worse than our the Pareto-optimal solutions.

For WMS dataset, similar results are recorded. Our proposed feature selection approach results have 0.78143 Hamming score with 14 features on the combination of CC-J48 algorithm. CBFS could only reach to 0.719 Hamming-score with 17 features and IG has Hamming-score to 0.71883 with 22 features. PCA has the worst results (see Table 13).

The execution times of all the algorithms are reported. Since our proposed feature selection method is a multi-objective evolutionary approach, the execution time is much higher than other state-of-the-art feature selection algorithms. Tables 14 and 15 present the execution time of state-of-the-art feature selection algorithms on MIR-Flickr and WMS datasets respectively.

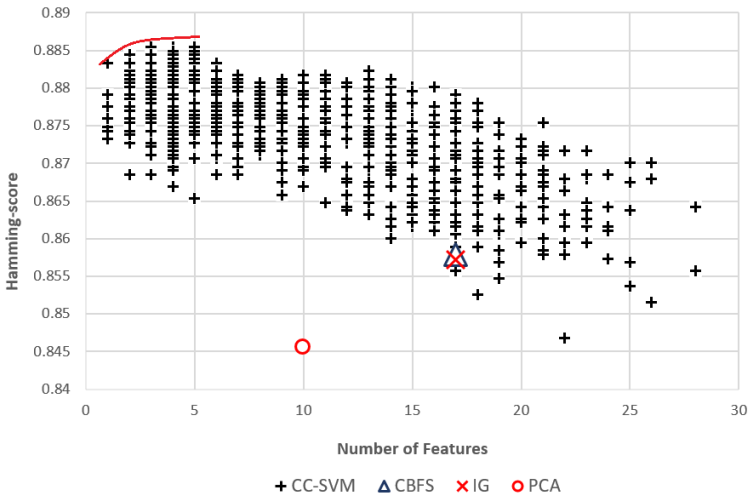

Fig. 15 The results of the CC-SVM multi-objective evolutionary algorithm and other state-of-the-art algorithms (CBFS, IG, PCA) on MIR-Flickr dataset. Red curve shows the Pareto optimal solutions of CC-SVM.

Table 12 Comparison of state-of-the-art feature selection algorithms on MIR-Flickr Dataset. The best results are given in bold numbers.

Multi-label Classification Algorithm	Base Classifier	Proposed Algorithm	# of Features After Proposed Algorithm Applied	CBFS (17)	IG (17)	PCA (10)
BR	LR	**0.86939**	12	0.86452	0.86428	0.86583
	SVM	**0.88494**	2	0.86570	0.86570	0.86778
	J48	**0.88865**	6	0.86335	0.86265	0.86409
CC	LR	**0.86532**	6	0.84726	0.84596	0.84543
	SVM	**0.88706**	7	0.85787	0.85726	0.84561
	J48	**0.88653**	6	0.84674	0.84574	0.84435
PS	LR	**0.82148**	15	-	-	-
	SVM	**0.86647**	11	0.84139	0.84165	0.84048
	J48	**0.86426**	15	0.81104	0.80615	0.80904
RAkEL	LR	**0.86439**	1	0.85783	0.85833	0.86313
	SVM	0.86265	1	**0.86543**	0.86496	0.86739
	J48	**0.86485**	1	0.81100	0.81178	0.81709

Table 13 Comparison of state-of-the-art feature algorithms on WMS Dataset. The best results are given in bold numbers.

Multi-label Classification Algorithm	Base Classifier	Proposed Algorithm	# Features After Proposed Algorithm Applied	CBFS (17)	IG (22)	PCA (10)
BR	LR	**0.84557**	32	0.75167	0.79200	0.71810
	SVM	**0.84937**	42	0.75200	0.79033	0.72238
	J48	**0.78481**	14	0.71700	0.72767	0.70143
CC	LR	0.77764	11	0.74033	**0.78833**	0.70238
	SVM	0.77595	13	0.75167	**0.78383**	0.72286
	J48	**0.78143**	14	0.71900	0.71883	0.67286
PS	LR	0.76709	45	0.74700	**0.78400**	0.71095
	SVM	**0.84473**	41	0.74700	0.78650	0.72810
	J48	**0.77004**	15	0.68700	0.71400	0.66000
RAkEL	LR	0.77131	21	0.74700	**0.78400**	0.71095
	SVM	0.76498	47	0.74700	**0.78650**	0.72810
	J48	**0.76708**	38	0.68700	0.71400	0.66000

Table 14 The execution time of algorithms on MIR-Flickr dataset

Algorithm	Execution Time (sec)
NSGA-II	1396.8
PCA	0.7804
CBFS	1.6866
IG	1.3012

Table 15 The execution time of algorithms on WMS dataset

Algorithm	Execution Time (sec)
NSGA-II	0.6482
PCA	0.0619
CBFS	0.4322
IG	0.4026

6 Conclusions and future work

In this paper, we propose a framework of multi-objective parallel evolutionary algorithms to both select the minimum number of multi-label image/video dataset features and provide the maximum prediction accuracy values. The feature selection process is implemented using the well-known NSGA-II algorithm and applied with twelve different combinations of various machine learning techniques. The experiments are carried out on two datasets (MIR-Flickr dataset and our WMS video recording dataset). The results of the experiments validate that the proposed algorithms improve the prediction accuracy of the results with a minimum number of features. We have observed that the Hamming score increases when the number of features is reduced during the multi-objective optimization process. The proposed algorithms succeed in obtaining Pareto-optimal solutions that have high prediction accuracy values with a minimum number of features.

The selection of the features being an open research problem, it is possible to improve the accuracy of the predictions with the new techniques proposed for the selection of the features. In addition, new algorithms enriched with deep learning techniques can be run on more powerful computing capabilities, increasing the number of generations, and exploring with diverse populations, which may produce better results. The use of deep auto-encoders for feature selection and the use of parallel scalable multi-objective optimization algorithms to select not only the minimum number of features for multi-tag image dataset, but also for an audio dataset and providing the maximum prediction accuracy values are our ongoing research.

Acknowledgment

This study is supported in part by NU Faculty development competitive research grants program, Nazarbayev University, Grant Number-110119FD4543 and in part by a research grant from TUBITAK (The Scientific and Technological Research Council of Turkey) with the grant number 114R082.

References

1. Alpaydin, E. (2014). Introduction to machine learning. MIT press.
2. Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919-926.
3. Srivastava, M. S., Joshi, M. N., & Gaur, M. (2014). A review paper on feature selection methodologies and their applications. IJCSNS, 14(5), 78.
4. Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In Proceedings of the 25th international conference on Machine learning (pp. 33-40).
5. Cerri, R., Basgalupp, M. P., Barros, R. C., & de Carvalho, A. C. (2019). Inducing hierarchical multi-label classification rules with genetic algorithms. Applied Soft Computing, 77, 584-604.
6. Gargiulo, F., Silvestri, S., Ciampi, M., & De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. Applied Soft Computing, 79, 125-138.

7. Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation, 6(2), 182-197.
8. Huiskes, M. J., & Lew, M. S. (2008). The MIR flickr retrieval evaluation. In Proceedings of the 1st ACM international conference on Multimedia information retrieval (pp. 39-43).
9. Costa, A. F., Traina, A. J. M., & Traina Jr, C. (2014). MFS-Map: efficient context and content combination to annotate images. In Proceedings of the 29th Annual ACM Symposium on Applied Computing (pp. 945-950).
10. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.
11. Yin, J., Tao, T., & Xu, J. (2015). A multi-label feature selection algorithm based on multi-objective optimization. In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
12. Zhang, Y., Gong, D. W., Sun, X. Y., & Guo, Y. N. (2017). A PSO-based multi-objective multi-label feature selection method in classification. Scientific reports, 7(1), 1-12.
13. Dokeroglu, T., Sevinc, E., Kucukyilmaz, T., & Cosar, A. (2019). A survey on new generation metaheuristic algorithms. Computers & Industrial Engineering, 137, 106040.
14. Vaishali, R., Sasikala, R., Ramasubbareddy, S., Remya, S., & Nalluri, S. (2017). Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. In 2017 International Conference on Computing Networking and Informatics (ICCNI) (pp. 1-5). IEEE.
15. Vignolo, L. D., Milone, D. H., & Scharcanski, J. (2013). Feature selection for face recognition based on multi-objective evolutionary wrappers. Expert Systems with Applications, 40(13), 5077-5084.
16. Labani, M., Moradi, P., Jalili, M., & Yu, X. (2017). An evolutionary based multi-objective filter approach for feature selection. In 2017 World Congress on Computing and Communication Technologies (WCCCT) (pp. 151-154). IEEE.
17. Zhang, P., Gao, W., & Liu, G. (2018). Feature selection considering weighted relevancy. Applied Intelligence, 48(12), 4615-4625.
18. Deniz, A., Kiziloz, H. E., Dokeroglu, T., & Cosar, A. (2017). Robust multiobjective evolutionary feature subset selection algorithm for binary classification using machine learning techniques. Neurocomputing, 241, 128-146.
19. Saroj, J. (2014). Multi-objective genetic algorithm approach to feature subset optimization. In Proc. of IEEE Intl Advance Computing Conf.(IACC) (pp. 544-548).
20. Hamdani, T. M., Won, J. M., Alimi, A. M., & Karray, F. (2007). Multi-objective feature selection with NSGA II. In International conference on adaptive and natural computing algorithms (pp. 240-247). Springer, Berlin, Heidelberg.
21. Khan, M. A., Ekbal, A., Menca, E. L., & Furnkranz, J. (2017). Multi-objective optimisation-based feature selection for multi-label classification. In International Conference on Applications of Natural Language to Information Systems (pp. 38-41). Springer.
22. Li, S., Wu, H., Wan, D., & Zhu, J. (2011). An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. Knowledge-Based Systems, 24(1), 40-48.
23. Gaspar-Cunha, A. (2010). Feature selection using multi-objective evolutionary algorithms: application to cardiac SPECT diagnosis. In Advances in bioinformatics (pp. 85-92). Springer, Berlin, Heidelberg.
24. Xue, B., Zhang, M., & Browne, W. N. (2012). Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE transactions on cybernetics, 43(6), 1656-1671.
25. Zhang, Y., Gong, D. W., & Cheng, J. (2015). Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE/ACM transactions on computational biology and bioinformatics, 14(1), 64-75.
26. Tangherloni, A., Spolaor, S., Cazzaniga, P., Besozzi, D., Rundo, L., Mauri, G., & Nobile, M. S. (2019). Biochemical parameter estimation vs. benchmark functions: a

comparative study of optimization performance and representation design. Applied Soft Computing, 81, 105494.
27. Nalluri, M. R., Kannan, K., Gao, X. Z., & Roy, D. S. (2019). Multiobjective hybrid monarch butterfly optimization for imbalanced disease classification problem. International Journal of Machine Learning and Cybernetics, 1-29.
28. Rundo, L., Tangherloni, A., Nobile, M. S., Militello, C., Besozzi, D., Mauri, G., & Cazzaniga, P. (2019). MedGA: a novel evolutionary method for image enhancement in medical imaging systems. Expert Systems with Applications, 119, 387-399.
29. Rundo, L., Tangherloni, A., Cazzaniga, P., Nobile, M. S., Russo, G., Gilardi, M. C., ... & Militello, C. (2019). A novel framework for MR image segmentation and quantification by using MedGA. Computer methods and programs in biomedicine, 176, 159-172.
30. Thabtah, F. A., Cowling, P., & Peng, Y. (2004). MMAC: A new multi-class, multi-label associative classification approach. In Fourth IEEE International Conference on Data Mining (ICDM'04) (pp. 217-224). IEEE.
31. Charte, F., del Jesus, M. J., & Rivera, A. J. (2016). Multilabel classification: problem analysis, metrics and techniques. Springer.
32. Read, J., Pfahringer, B., & Holmes, G. (2008). Multi-label classification using ensembles of pruned sets. In 2008 eighth IEEE international conference on data mining (pp. 995-1000). IEEE.
33. Tsoumakas, G., & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In European conference on machine learning (pp. 406-417). Springer, Berlin, Heidelberg.
34. Lobato, F. S., & Steffen, V. (2017). Multi-Objective Optimization Problem. In Multi-Objective Optimization Problems (pp. 9-23). Springer, Cham.
35. Zhou, A., Qu, B. Y., Li, H., Zhao, S. Z., Suganthan, P. N., & Zhang, Q. (2011). Multiobjective evolutionary algorithms: A survey of the state of the art. Swarm and Evolutionary Computation, 1(1), 32-49.
36. Stadler, W. (1979). A survey of multicriteria optimization or the vector maximum problem, part I: 17761960. Journal of Optimization Theory and Applications, 29(1), 1-52.
37. Miettinen, K. (2012). Nonlinear multiobjective optimization (Vol. 12). Springer Science & Business Media.
38. Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. Machine learning, 85(3), 333.
39. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
40. Zeng, Z. Q., Yu, H. B., Xu, H. R., Xie, Y. Q., & Gao, J. (2008). Fast training support vector machines using parallel sequential minimal optimization. In 2008 3rd international conference on intelligent system and knowledge engineering (Vol. 1, pp. 997-1001). IEEE.
41. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
42. Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6).
43. Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22).
44. Dokeroglu, T., & Sevinc, E. (2019). Evolutionary parallel extreme learning machines for the data classification problem. Computers & Industrial Engineering, 130, 237-249.
45. Cantu-Paz, E. (1998). A survey of parallel genetic algorithms. Calculateurs paralleles, reseaux et systems repartis, 10(2), 141-171.
46. Hadka, D. (2014). MOEA framework user guide.
47. Costa, A. F., Humpire-Mamani, G., & Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. In 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images (pp. 39-46). IEEE.
48. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

49. Bradski, G., & Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc.".
50. Read, J., Reutemann, P., Pfahringer, B., & Holmes, G. (2016). Meka: a multi-label/multi-target extension to weka. The Journal of Machine Learning Research, 17(1), 667-671.
51. Tan, Q., Yu, G., Domeniconi, C., Wang, J., & Zhang, Z. (2018). Incomplete multi-view weak-label learning. In IJCAI (pp. 2703-2709).
52. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559-572.
53. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.