The Knowledge Engineering Review (2023), **38**, e1, pp. 1–43
doi:10.1017/S0269888923000024

The Knowledge Engineering Review
CAMBRIDGE UNIVERSITY PRESS

REVIEW

A survey of evolutionary algorithms for supervised ensemble learning

Henry E. L. Cagnini[1], Silvia C. N. Das Dôres[1], Alex A. Freitas[2] and Rodrigo C. Barros[1]

[1]School of Technology, PUCRS, Avenida Ipiranga 6681, Porto Alegre, RS 90619-900, Brazil;
e-mails: henry.cagnini@edu.pucrs.br, silvia.dores@acad.pucrs.br, rodrigo.barros@pucrs.br
[2]Computing School, University of Kent, Giles Ln, Canterbury CT2 7NZ, UK;
e-mail: a.a.freitas@kent.ac.uk

Received: 27 October 2021; **Revised:** 11 January 2023; **Accepted:** 11 January 2023

Abstract

This paper presents a comprehensive review of evolutionary algorithms that learn an ensemble of predictive models for supervised machine learning (classification and regression). We propose a detailed four-level taxonomy of studies in this area. The first level of the taxonomy categorizes studies based on which stage of the ensemble learning process is addressed by the evolutionary algorithm: the generation of base models, model selection, or the integration of outputs. The next three levels of the taxonomy further categorize studies based on methods used to address each stage. In addition, we categorize studies according to the main types of objectives optimized by the evolutionary algorithm, the type of base learner used and the type of evolutionary algorithm used. We also discuss controversial topics, like the pros and cons of the selection stage of ensemble learning, and the need for using a diversity measure for the ensemble's members in the fitness function. Finally, as conclusions, we summarize our findings about patterns in the frequency of use of different methods and suggest several new research directions for evolutionary ensemble learning.

1. Introduction

Supervised ensemble learning—sometimes referred to as a mixture of experts, classifier ensembles, or multiple classifier system (Saleh *et al.,* 2016; Milliken *et al.,* 2016) is a paradigm within the machine learning area concerned with integrating multiple base supervised learners in order to produce better predictive models than simply learning a single strong model. An ensemble typically performs its predictions by using a voting mechanism (e.g., majority voting) that computes the mean or the mode of the predictions output by the ensemble's members (base learners). Ensemble learning methods have won several academic and industrial machine learning competitions (Sagi & Rokach, 2018), and such methods have been extensively deployed in real-world AI applications (Oza & Tumer, 2008; Tabassum & Ahmed, 2016).

Ensembles have several advantages over a single learner: (i) it is usually computationally cheaper to integrate a set of simple, weak models than to induce a single robust, complex model (Krithikaa & Mallipeddi, 2016); (ii) ensembles composed by classifiers that are, in turn, only slightly better than random guessing, can still present predictive performance comparable to a strong single classifier (Freund & Schapire, 1995; Liu *et al.,* 2009; Jackowski *et al.,* 2014); (iii) different base learners can be specialized in different regions of the input space, making their consensus more flexible and effective when dealing with complex problems (Freund & Schapire, 1995). Indeed, there is both theoretical and empirical evidence demonstrating that a good ensemble can be obtained by combining individual models that

Cite this article: H. E. L. Cagnini, S. C. N. Das Dôres, A. A. Freitas and R. C. Barros. A survey of evolutionary algorithms for supervised ensemble learning. *The Knowledge Engineering Review* 38(e1): 1–43. https://doi.org/10.1017/S0269888923000024

make distinct errors (e.g., errors on different parts of the input space) (Kim & Cho, 2008a; Hansen & Salamon, 1990; Krogh & Vedelsby, 1995; Opitz & Maclin, 1999; Hashem, 1997).

Ensemble learning comprises three distinct stages, whose names vary in the literature: generation, selection, and integration (Castro & Von Zuben, 2011; Britto *et al.,* 2014; Lima & Ludermir, 2015), which is the most common naming system, and the one that will be used in this paper; pre-gate, ensemble-member, and post-gate (Debie *et al.,* 2016); or generation, pruning, and fusion (Parhizkar & Abadi, 2015a). The three-step generation of ensembles can be reduced to a hypothesis-search problem in combinatorial spaces, and so it is often approached by a variety of heuristic approaches, such as boosting (Freund & Schapire, 1995), bagging (Breiman, 1996), and Evolutionary Algorithms (EAs) (Lima & Ludermir, 2015; Lacy *et al.,* 2015b).

EAs have several advantages for ensemble learning, such as: performing a global search, which is less likely to get trapped into local optima than greedy search methods (Xavier-Júnior *et al.,* 2018); being easily adapted for multi-objective optimization (Deb, 2011); and dealing with multiple solutions in parallel, due to their population-based nature, for example Hauschild and Pelikan (2011), Kumar *et al.* (2010). Hence, many EAs for supervised ensemble learning have been proposed in the literature in the past few years.

Several application domains have benefited from EA-based ensemble learning algorithms, including for example wind speed prediction (Woon & Kramer, 2016; Zhang *et al.,* 2017a), cancer detection (Krawczyk *et al.,* 2013; Krawczyk & Woźniak, 2014; Krawczyk & Schaefer, 2014; Krawczyk *et al.,* 2015, 2016; Singh *et al.,* 2016), noise bypass detection in vehicles (Redel-Macías *et al.,* 2013), stock market prediction (Chen *et al.,* 2007; Mabu *et al.,* 2014, 2015), and microarray data classification (Park & Cho, 2003b, 2003a; Kim & Cho, 2005, 2008a; Liu *et al.,* 2007, 2014a, 2015; Chen & Zhao, 2008; Rapakoulia *et al.,* 2014; Kim & Cho, 2015; Ali & Majid, 2015; Saha *et al.,* 2016), to name just a few.

This paper presents a survey of EAs for supervised ensemble learning. Our main contribution is to properly identify, categorize, and evaluate the available research studies in this area. This survey is aimed toward researchers on evolutionary algorithms and/or ensemble learning algorithms.

As related work, several surveys have been published on ensemble studies from different perspectives. Regarding specifically EAs for supervised ensemble learning, Yao and Islam (2008) present a review of EAs for designing ensembles, but they focus only on artificial neural networks (ANNs) as the base learners to be combined. Sagi and Rokach (2018), as well as Dietterich (2000) present a general review of ensemble learning studies, based on traditional non-evolutionary methods. Rokach (2010), Kotsiantis (2014), and Tabassum and Ahmed (2016) review ensembles designed only for classification tasks. Similarly, Mendes-Moreira *et al.* (2012) and Vega-Pons and Ruiz-Shulcloper (2011) review ensemble methods for regression and clustering tasks only, respectively. There are also papers on specific domain applications of ensembles, for example Athar *et al.* (2017), which reviews ensemble methods for sentiment analysis; Gomes *et al.* (2017) and Krawczyk *et al.* (2017) also review ensemble learning for data stream classification and regression.

Despite the relevant contributions of the previously cited literature, this work is to the best of our knowledge the first review to focus on general-application EAs for supervised ensemble learning in a comprehensive fashion. In particular, we highlight the following contributions: (i) we provide a general overview of EAs for supervised ensemble learning, not exclusively focusing on any specific EA or any given type of supervised model, but presenting an in-depth analysis of the different algorithms proposed for each stage of ensemble learning, with their respective advantages and pitfalls; and (ii) we provide a detailed taxonomy to properly categorize supervised evolutionary ensembles, helping the reader to filter the literature and understand the possibilities when designing EAs for this task. Note that reviewing EAs for ensemble learning in unsupervised settings (e.g., the clustering task) is out of the scope of this paper.

The rest of this paper is organized as follows. Section 2 briefly reviews the most well-known types of ensemble learning methods. Section 3 presents our novel taxonomy to categorize EAs designed for supervised ensemble learning. Sections 4, 5, and 6 review the EAs employed for the three stages of ensemble learning: generation, selection, and integration. In the next sections, we focus on broader aspects of EAs for ensemble learning that are not specific to any single stage, as follows. Section 7

details types of fitness functions often used by EAs for ensemble learning. Section 8 summarizes the types of EAs used for ensemble learning. Section 9 points to the most common base learners within the EAs reviewed in this survey. Section 10 describes the complexity of evolutionary algorithms when applied to ensemble learning. Finally, in Section 11 we summarize our findings by identifying patterns in the frequency of use of different methods across the surveyed EAs, and identify future trends and interesting research directions in this area.

2. Ensemble learning

There are three main motivations to combine multiple learners (Dietterich, 2000): representational, statistical, and computational. The representational motivation is that combining multiple base learners may provide better predictive performance than a single strong learner. For example, the generalization ability of a neural network can be improved by using it as base learner within an ensemble (Chen & Yao, 2006). In theory, no base learner will have the best predictive performance for all problems, as stated by the *No Free Lunch* theorem (Wolpert & Macready, 1997); and in practice, selecting the best learner for any given dataset is a very difficult problem (Fatima *et al.,* 2013; Kordík *et al.,* 2018), which can be addressed by integrating several good learners into an ensemble.

The statistical motivation is to avoid poor performance by averaging the outputs of many base learners. While averaging the output of multiple base learners may not produce the overall best output, it is also unlikely that it will produce the worst possible output (Hernández *et al.,* 2015). This is particularly the case for data with few data points, so overfitting is more likely.

Finally, the computational motivation is that some algorithms require several runs with distinct initializations in order to avoid falling into bad local minima. Gradient descent, for example, often requires several runs and further evaluation on a validation set in order to avoid being trapped into local minima. Thus, it seems reasonable to integrate these already-trained intermediate models into an ensemble, stabilizing and improving the system's overall performance (Tsakonas & Gabrys, 2013; de Lima & Ludermir, 2014).

Ensemble learning became popular during the 1990's (de Lima & Ludermir, 2014), with some of the most important work arising around that time: stacking in 1992 (Wolpert, 1992); boosting in 1995 and 1996 (Freund & Schapire, 1995, 1996; Schapire, 1999); bagging in 1996 (Breiman, 1996); and random forests in the early 2000's (Breiman, 2001). We call these methods *traditional* to differentiate them from EA-based ensembles, though they are also referred to as *preprocessing-based ensemble methods* in the EA literature (Krawczyk *et al.,* 2016). We briefly review them next.

2.1 Boosting

Boosting refers to the technique of continuously enhancing the predictive performance of a weak learner (Krawczyk *et al.,* 2016). We present here the popular AdaBoost algorithm, proposed by Freund and Schapire (1995). Given a set of predictive attributes $\mathbf{X}$ and a set of class labels Y, $y \in Y = \{-1, +1\}$, in its first iteration AdaBoost assigns equal importance (weights) to each instance in the training set, $D_1(i) = 1/N$, $i = 1, \ldots, N$, with N as the number of instances. For a given number of iterations G, AdaBoost trains a weak classifier based on the distribution D_g and then computes its error $\epsilon_g = \mathrm{P}_{i \sim D_g}[h_g(x_i) \neq y_i]$. Instances that are harder to classify will have their weights increased, so it becomes more rewarding to the model to classify them correctly.

Candidate algorithms for boosting must support assignment of weights for instances. If this is not possible, a set of instances can be sampled from D_g and supplied to the gth learner. Although boosting usually improves the predictive performance of a weak classifier, its performance suffers when faced with noisy instances, since failing to correctly classify those instances will iteratively improve their importance and hence lead the learner to overfitting (Lacy *et al.,* 2015b).

2.2 Bagging

Bootstrap **agg**regat**ing**, or simply *bagging*, aims at reducing training instability when a learner is faced with a given data distribution (Breiman, 1996). It consists of generating B subsets of size N from the original training distribution $D(i) = 1/N$, $i = 1, \ldots, N$ with replacement, causing some instances to be present in more than one subset. As a result, some base learners have a tendency to favor such instances, having more opportunities to correctly predict their values. The predictions of all trained B learners are combined by computing their mean (regression task) or mode (classification task).

By sampling different subsets of instances for different classifiers, bagging implicitly injects diversity within the ensemble (Lacy *et al.,* 2015b), whereas boosting explicitly does this by weighting the data distribution to focus the base learners' attention into more difficult instances (Gu & Jin, 2014; Lacy *et al.,* 2015b).

2.3 Stacking

Compared to bagging and boosting, *stacked generalization* or simply *stacking* (Wolpert, 1992) is a more flexible strategy for ensemble learning. The user can choose one or more types of base learners to be used in the ensemble (e.g., using only decision trees, or mixing them with ANNs Shunmugapriya & Kanmani, 2013). Then, each base learner will output a prediction, and all learners' predictions will be combined by a meta-learner (which can also be chosen) to produce a single output. Popular traditional meta-learners for stacking include linear regression (for regression) and logistic regression (for classification). In Section 6.1 we review evolutionary algorithms used for learning logistic regression and linear regression algorithms' weights. Stacking often improves the overall predictive performance of ensembles, making it a popular method (Shunmugapriya & Kanmani, 2013; Milliken *et al.,* 2016).

2.4 Random forests

Random forests, proposed by Breiman (2001), is a type of ensemble learning method where both the base learner and data sampling are predetermined: decision trees and random sampling of both instances and attributes. The training process for the original random forests algorithm (Breiman, 2001) is described as follows. First, the algorithm randomly samples with replacement B subsets of training instances, one for building each of B decision trees that will compose the ensemble. For each inner node within a decision tree, the algorithm first randomly samples without replacement a subset of m attributes, and then it selects, among those attributes, the one that minimizes the local class impurity for that node. In this context, purity is the ratio of instances from difference classes that follow the same tree branch; hence, maximum purity in a node means that all instances reaching said node belong to the same class. This procedure is applied to each inner node in the current decision tree within the ensemble, and it is repeated until the tree achieves maximum class purity for all leaf nodes.

Random forests sometimes perform better than boosting methods, while being resilient to outliers and noise, faster to train than bagging and boosting methods (depending on the respective base learner), and being easily parallelized. However, it can require very many decision trees to provide an acceptable predictive performance, depending on the dataset at hand. Table 1 presents a brief overview of the main characteristics of the above traditional methods.

3. Ensemble learning with evolutionary algorithms

In general, Evolutionary Algorithms (EAs) are robust optimization methods that perform a global search in the space of candidate solutions. EAs are simple to implement, requiring little domain knowledge and can produce several good solutions to the same problem due to their population-based nature (Galea *et al.,* 2004). In particular, EAs seem to be naturally suited for ensemble learning,

Table 1. *Traditional ensemble learning methods compared. Adapted from Ma et al. (2015)*

Algorithm	Sampling	Base learner	Integration strategy
Bagging (Breiman, 1996)	instance	Unstable learner trained over re-sampled instance subsets	Majority voting
Boosting (Freund & Schapire, 1995)	instance	Weak learner re-weighted at every iteration	Weighted majority voting
Stacking (Wolpert, 1992)	None	Any	Meta-model
Random forests (Breiman, 2001)	instance; attribute	Decision trees	Majority voting

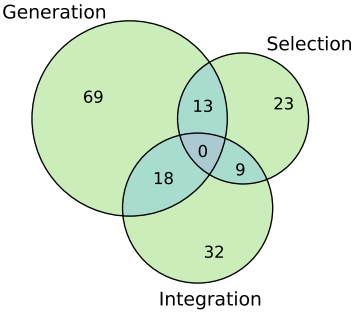

Figure 1. *Work summarized by the ensemble learning stage that EAs are employed. While generation is more popular than selection and integration combined, none of the surveyed studies employed EAs in all stages.*

given their capability of producing a set of solutions that can be readily integrated into an ensemble (Duell *et al.,* 2006; Lacy *et al.,* 2015b). EAs also support multi-objective optimization (based e.g., on Pareto dominance), allowing the generation of solutions that cover distinct aspects of the input space (Lacy *et al.,* 2015b) and removing the need to manually optimize some hyper-parameters (e.g., the base learner's hyper-parameters, the number of ensemble members, etc.). However, EAs will likely increase the computational cost of ensemble learning, due to its robust global-search behavior that usually considers many tens or hundreds of possible solutions at each iteration (generation). Nonetheless, parallelization is an option to mitigate such problem (Lima & Ludermir, 2015).

Since ensemble learning is composed of at least three main optimization steps (generation, selection, and integration), each one with many tasks, there is plenty of room to employ EAs (Fernández *et al.,* 2016b). In the literature on EA-based ensembles for supervised learning, the most common approach is to optimize a single step, though some studies go as far as optimizing two of them (e.g., Chen & Zhao, 2008; Ojha *et al.,* 2017). Figure 1 summarizes how many studies were dedicated to each of the three stages.

In this work, we provide a taxonomy to categorize the EA-based approaches for supervised ensemble learning (Figure 2). All surveyed studies are focused on supervised problems, that is, no unsupervised approach is reviewed.

We divide the surveyed studies according to the well-established main stages of supervised ensemble learning: generation, selection, and integration. We use these three stages at the top level of our proposed taxonomy because in principle major decisions about the design of the EA (e.g., which individual representation to use, which fitness function to use) are entirely dependent on the type of ensemble learning stage addressed by the EA. The approaches most often used in each stage are presented at the second

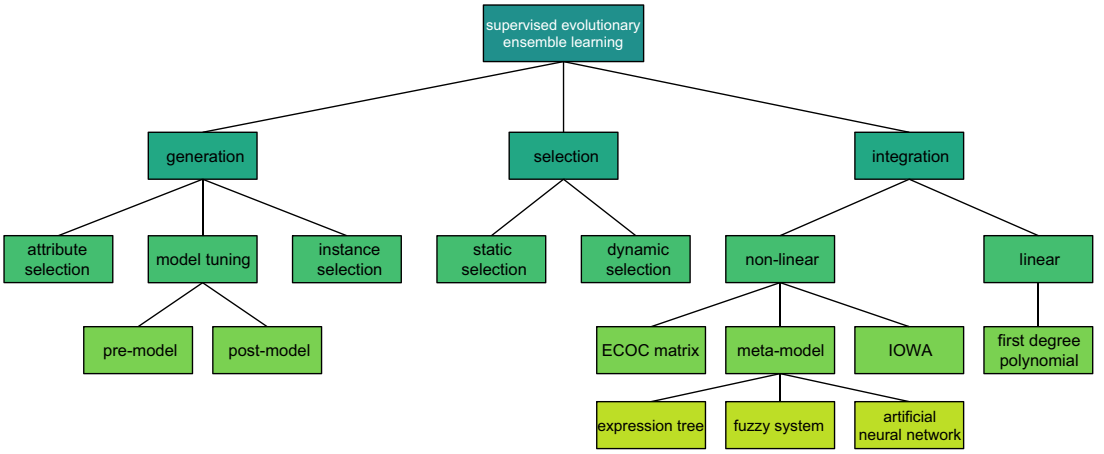

Figure 2. *The proposed taxonomy for EAs employed in ensemble learning.*

level of the taxonomy. For example, attribute selection, model tuning, and instance selection are the three most common approaches for the generation stage. Further divisions in the taxonomy are presented at the next levels, whenever it is the case.

Note that taxonomies vary depending on the aspect being analyzed—for example Gu (Gu & Jin, 2014) is concerned with the generation stage and hence proposes a taxonomy exclusively for that step. To the best of our knowledge, our taxonomy is one of the broadest with regard to EAs for ensemble learning, with the closest reference being the one proposed by Cruz *et al.* (2018). While the description of generation and selection stages in Cruz *et al.* (2018) is identical to ours, we are more specific regarding the strategies for the integration stage. In addition, while the authors propose a two-level taxonomy, we present a more detailed and thorough four-level taxonomy.

3.1 Methodology to collect and analyze papers in this survey

The main objective of this work is to identify and evaluate existing approaches that apply evolutionary algorithms for learning ensembles of predictive models for supervised machine learning. The objective is expressed from the research questions presented in Table 2. These questions aim to analyze the relevant work, both in the context of evolutionary algorithms used, and the characteristics of ensembles that are optimized. The sections where these questions are addressed are also shown in the table.

3.1.1 Search strategy

Based on the main objective, we select keywords that are likely to be present in most of the work that proposes evolutionary algorithms for ensemble learning; from these keywords we compose a search string. Synonyms of each term were incorporated using the Boolean operator *OR*, whereas the Boolean operator *AND* was used to link the terms. The generic search string derived is

```
‘ensemble’ AND
(‘classification’ OR ‘classifier’ OR ‘classifiers’) OR
(‘regression’ OR ‘regressor’ OR ‘regressors’) AND
(‘evolutionary’ OR ‘evolution’)
```

Table 2. *Research questions of this survey*

ID	Research question	Description	Addressed in
RQ1	What are the existing work that apply evolutionary algorithms for learning ensembles for supervised machine learning tasks?	General question that aims to identify existing work that apply evolutionary algorithms in the context of ensemble learning	Throughout this survey
RQ2	What are the evolutionary algorithms used to learn the ensembles?	Aims to identify which evolutionary algorithms are applied for ensemble learning	Section 8
RQ3	What stages of ensemble learning are addressed by the evolutionary algorithm?	Aims to categorize the approaches according to the ensemble optimization step (generation, selection or integration)	Sections 4, 5, and 6
RQ4	Which objective functions are optimized by the evolutionary algorithm?	Since fitness function is an essential component of EAs, and given the complexity of the ensembles where several objectives can be optimized, this question aims to analyze how these functions are employed in ensemble learning	Section 7
RQ5	What are the base learners used?	Finally, this survey aims to analyze the relevant works from the point of view of the base learners that are used to compose the ensembles	Section 9

In addition to the search string, we define the search engines. Thus, reviewed papers of this survey were searched in the following repositories: Scopus[1], Science Direct[2], IEEE Xplore[3], and ACM Digital Library[4]. Figure 3 shows the search strings as used in each search engine.

3.1.2 Study selection criteria

We adopted the following criteria for including studies in this survey: (i) papers that present a new evolutionary strategy for ensemble learning in supervised machine learning; (ii) papers that present a minimum detail of the proposed solution, including: type of EA used, fitness function used, ensemble stage optimized, base learners used; and (iii) papers that present an experimental evaluation of the proposed solution. We also use exclusion criteria, which are: (i) unavailability: paper not available in any online repository, or papers available only under payment; (ii) wrong topic: on further review, papers that did not cover the surveyed topic; and (iii) papers that are not written in English.

[1] Available at https://www.scopus.com/home.uri.
[2] Available at http://www.sciencedirect.com.
[3] Available at http://ieeexplore.ieee.org/Xplore/home.jsp.
[4] Available at http://dl.acm.org.

(a)

```
TITLE-ABS-KEY("ensemble") AND (
    (
        TITLE-ABS-KEY("classification") OR
            TITLE-ABS-KEY("classifier") OR
            TITLE-ABS-KEY("classifiers")
    ) OR (
        TITLE-ABS-KEY("regression") OR
        TITLE-ABS-KEY("regressor") OR
        TITLE-ABS-KEY("regressors")
    )
) AND (
    TITLE-ABS-KEY("evolutionary") OR
    TITLE-ABS-KEY("evolution")
)
```

Scopus

(b)

```
"ensemble" AND (
    (
        "classification" OR
        "classifier" OR
        "classifiers"
    ) OR (
        "regression" OR
        "regressor" OR
        "regressors"
    )
) AND (
    "evolutionary" OR
    "evolution"
)
```

ACM

(c)

```
"Abstract":ensemble AND (
    (
        "Abstract":classification OR
        "Abstract":classifier OR
        "Abstract":classifiers
    ) OR (
        "Abstract":regression OR
        "Abstract":regressor OR
        "Abstract":regressors
    )
) AND (
    "Abstract":evolutionary OR
    "Abstract":evolution
)
```

IEEE Xplore

(d)

```
title-abstr-key("ensemble") AND (
    (
        title-abstr-key("classification") OR
        title-abstr-key("classifier") OR
        title-abstr-key("classifiers")
    ) OR (
        title-abstr-key("regression") OR
        title-abstr-key("regressor") OR
        title-abstr-key("regressors")
    )
) AND (
        title-abstr-key("evolutionary") OR
        title-abstr-key("evolution")
)
```

ScienceDirect

Figure 3. *Search strings as used in search engines.*

3.1.3 Study selection procedures

In the selection process, the search string was applied to the title, abstract, and keywords of searched papers. Scopus was the first repository searched, since it has the largest database. Eight hundred and two (802) papers matched the keywords. All papers had their abstracts reviewed, and from their analysis the ones deemed relevant (366) were carried on to the next stage of the reviewing process, as shown in column Relevant of Table 3.

We proceed the search with ACM Digital Library, IEEE Xplore, and Science Direct, again reviewing abstracts and selecting papers based on their relevance to this survey. Since Scopus is the largest database, some papers present in the remaining repositories were also present in Scopus. For this reason, column Already in Scopus of Table 3 counts the number of papers found in other repositories that already had their abstracts reviewed when we collected papers from Scopus. Among the remaining databases, we found 108 unique papers, not present in Scopus; from these, 38 were deemed relevant for further review.

Across all searched databases, 404 papers were deemed relevant for the survey, taking into consideration the description of their abstracts. From these, 50 were unavailable, either because (i) the document was not found in their host websites, or it was behind a paywall; or (ii) on further review of the paper, the topic addressed was not exactly the one we were interested (43 papers), as discussed in the beginning of Section 3.1.2. This reduced the number of relevant papers to 311.

From the remaining 311 papers, 164 were fully reviewed and included in the survey, with the remaining 147 not included nor reviewed. While we could have reviewed the latter group, we did not because we applied a truncation factor: that is, the rate at which we were detecting new concepts in papers was not justified by the amount of papers needed to reach these novel ideas. The papers that were not reviewed did not have any characteristic that made them less attractive than the ones reviewed, and we do not discriminate based on vehicle of publication, type of publication (conference or journal paper), date, number of citations, etc. An overall summary of the papers is presented in Figure 4.

Papers were reviewed in chronological order: the ones closest to the date of the reviewing process were added first, and the ones that had been already been published, reviewed last. From the papers that

Table 3. *Papers that matched the search strings shown in Figure 3. In this stage all papers had their abstracts reviewed. From this initial analysis, not all papers were deemed relevant to the scope of this survey. Already in Scopus column denote papers that were already present at the Scopus database*

Repository	Relevant	Irrelevant	Already in Scopus	Total
Scopus	366	436	–	802
ACM digital library	16	11	34	61
IEEE Xplore	13	8	127	148
Science direct	9	51	90	150
Total	**404**	506	251	1161

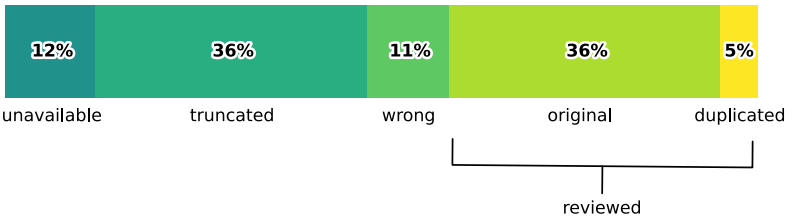

Figure 4. *From the* 404 *papers selected for review,* 164 *were added to the survey. Among these,* 20 *were duplicated (e.g., expanded work), and* 144 *original work.*

made to the survey, 20 were duplicated and fell into one of the following categories: (i) the algorithms were published in conferences and had expanded versions in journals; (ii) they had different application domains, but the algorithm was the same; or (iii) slightly different implementations (for example, changing the number of layers and/or activations in a neural network).

3.1.4 Data extraction strategy

After selecting the 164 works to compose the survey, the extraction and analysis of the data was made through peer review, where at least two researchers evaluated each work. The data were structured in a spreadsheet according to its meta-data (catalog information) and the characteristics of the work, according to the research questions that we aimed to answer.

We make available two supplementary material to this paper. The first is a repository of source code, hosted at https://github.com/henryzord/eael, with metadata used to generate figures and tables in this paper. The other is a master table, made available as a website, and hosted at https://henryzord.github.io/eael, listing individual information on surveyed work.

4. The generation stage of ensemble learning

In this stage the algorithm generates a pool of trained models. Those models may come from: (i) different paradigms (e.g., Naïve Bayes, support vector machines (SVMs), and Neural Networks Peimankar *et al.,* 2017); or (ii) be from the same paradigm, but still present some differences, such as ANNs with different topologies and/or activation functions (Zhang *et al.,* 2017a).

The main objective in this stage is to generate a pool of both accurate and diverse base learners. Base learners must be diverse in order to provide source material for the selection and integration steps to work with. A diverse pool of base learners has more chances to commit errors in different data instances, thus correctly predicting more instances (Pagano *et al.,* 2012).

An example of an ensemble algorithm that focuses on the generation phase is random forests (Breiman, 2001; Trawiński *et al.,* 2013), considering that it selects distinct subsets of both attributes and instances for building different decision trees, resulting in an ensemble of trees that is more robust than a single decision tree.

We have identified three ways of generating pools of learners that are both diverse and accurate: (i) providing distinct training sets for each base learner (instance selection); (ii) providing the same training set for all learners but with distinct sets of attributes (attribute selection); and (iii) optimizing the model by modifying the hyper-parameters and/or the structure of the base learners.

4.1 Instance selection

Instance selection, also known as prototype selection or data randomization (Vluymans *et al.,* 2016; Albukhanajer *et al.,* 2017) consists of providing different (not necessarily disjoint) subsets of training instances for different base learners (Almeida and Galvão, 2016; Rosales-Pérez *et al.,* 2017). This approach is well suited for homogeneous sets of base learners which are sensitive to changes in the instance distribution (e.g., decision trees Hernández *et al.,* 2015).

Instance selection can also be used to reduce training time by finding a subset of representative instances for each class (Almeida & Galvão, 2016; Rosales-Pérez *et al.,* 2017). This is also beneficial for problems with high class imbalance, given that re-sampling instances with replacement makes it possible to simulate a uniform distribution among classes. Indeed that is one of the capabilities of the traditional bagging algorithm (Breiman, 1996). Thus, 'bagged' EAs are likely to have the same benefits as 'bagged': improved noise tolerance and reduced overfitting risk (Vluymans *et al.,* 2016). A method for selecting instances is needed since random sampling can lead to information loss and poor model generalization (Karakatič *et al.,* 2015). By using an EA, both the tasks of undersampling the majority class and oversampling the minority class are possible in parallel.

This section mainly focuses on instance selection techniques, since instance generation is more scarce. While the former performs a selection of instances from the original data, the latter can create new artificial instances, thus easing the adjustment of decision boundaries of models, at the expense of being more prone to overfitting (Vluymans *et al.,* 2016). Only one work uses a hybrid selection-and-generation strategy, namely (Vluymans *et al.,* 2016). In this work, seeking to address the class imbalance problem, a Steady State Memetic Algorithm (SSMA) is used for selecting instances from the training set, composing several individuals (i.e., subsets of instances). Once the SSMA optimization ends, a portion of its (fittest) individuals will be then fed to a Scale Factor Local Search in Differential Evolution (SFLSDE), that will improve the quality of instance subsets by generating new instances. Both evolutionary algorithms use a measure of predictive performance as the fitness function. Finally, the (fittest) individuals from SFLSDE are used by 1-NN classifiers, integrated by means of weighted voting (not evolutionary induced).

Although instance selection is present in work tackling the class imbalance problem (e.g., Cao *et al.,* 2013a; Galar *et al.,* 2013; Vluymans *et al.,* 2016), it can lead to overfitting, or having subsets where one class has many more instances than other classes, if an inadequate objective function (such as accuracy) is used. An approach to avoid overfitting is to assign different misclassification costs to different classes. Typical cost-sensitive learning techniques do not directly modify the data distribution, but rather take misclassification costs into account during model construction (Cao *et al.,* 2013a).

Instance selection can be divided into wrapper and filter methods. In our literature review, wrapper methods were more common than filter methods. Only the work of Almeida and Galvão (2016) uses a filter approach, where a GA is used to optimize the number of groups in a k-means algorithm. Due to the tendency of k-means in finding hyperspherical clusters, the objective is to find evenly distributed groups of instances. One classifier is trained for each group, and the quality of classifiers is assessed by the weighted combination of training accuracy, validation accuracy, and distribution of classes within groups. In the prediction phase, unknown instances will be assigned to their most similar cluster, based on the Euclidean distance between the unknown-class test instance and the cluster's instances.

For wrapper methods, usually the set of selected instances is encoded as either a binary or real-valued chromosome of N positions (the number of instances). In the binary case, each bit encodes whether an instance is present or absent in the solution encoded by the current individual. In a real-valued case, each gene encodes the probability that the respective instance will be present in that solution. To address class imbalance, in Galar *et al.* (2013) only majority class instances are encoded in a binary string—minority class instances are always selected.

It is also possible to perform instance selection together with other methods. In Rosales-Pérez *et al.* (2017), the multi-objective problem consists in optimizing both a SVM's hyper-parameters and the instance set to be used for training each model. This combined strategy is better for SVMs than simply selecting training subsets, since SVMs are robust to small changes in data distribution (Batista *et al.,* 2017). Coupling two tasks at once also fits well with weight optimization: in Krawczyk *et al.* (2016), evolutionary undersampling and boosting are used in a C4.5 decision tree classifier to iteratively optimize its performance in grading breast cancer malignancy.

Olvera-López provides an extensive survey of both evolutionary and non-evolutionary instance-selection methods proposed until 2010 (Olvera-López *et al.,* 2010).

4.2 Attribute selection

Attribute selection, also called feature selection or variable subset selection (Sikdar *et al.,* 2012), offers distinct subsets of attributes to different base learners in order to induce diversity among base models. By removing irrelevant and redundant attributes from the data, attribute selection can improve the performance of base learners (Sikdar *et al.,* 2014b). Reducing the number of attributes also reduces the complexity of learned base models and may improve the efficiency of the ensemble system.

Attribute selection also performs dimensionality reduction and is an efficient approach to build ensembles of base learners (Liu *et al.,* 2009). There is no need to provide disjoint sets of attributes to different learners. The base learners must be sensitive to modifications in the data distribution. SVMs, for example, were reported to be little affected by attribute selection (Vaiciukynas *et al.,* 2014).

Wrapper methods are by far the most common type of EAs for attribute selection. It has been noted that there is a direct link between high-quality attribute subsets and a high-quality pool of base learners (Mehdiyev *et al.,* 2015). Filter approaches break this link, evaluating the quality of an attribute set in a way independent from the overall base learner pool (Mehdiyev *et al.,* 2015).

A wrapper method provides a reduced subset of attributes to a learning algorithm, and then the predictive accuracy of the model trained with those attributes is used as a measure of the quality of the selected attributes. The random subspace method, for example, is a traditional approach for wrapping algorithms that randomly selects different attribute subsets for different base learners. Although this method is usually much faster than EAs, its performance is sensitive to the number of attributes and ensemble size (Liu *et al.,* 2009). By contrast, EAs can improve stability and provide more accurate ensembles (Liu *et al.,* 2009). Other examples of traditional methods include sequential forward selection, sequential backward selection, beam search, etc. (Mehdiyev *et al.,* 2015).

However, filter methods are still usually preferred for some application domains, such as microarray data, where the number of attributes far surpasses the number of instances, rendering a wrapper approach inefficient. In Kim and Cho (2008a), base learners are coupled with filters that perform attribute selection. Although the authors do not use training time as an objective in the EA, the reported execution time of a single run of the GA is between 15 seconds to 3 minutes, much faster than an exhaustive search, that could take as long as one hour (for sets of 24 attributes), or one year (for sets of 42 attributes), in a dataset of 4026 attributes. Their proposed algorithm is also capable of outperforming other EA-based ensembles for two microarray datasets. In another study, genetic algorithms (GAs) with error rate as fitness function were shown to be capable of outperforming greedy wrapper methods in terms of ensemble accuracy (Mehdiyev *et al.,* 2015).

Two concepts relevant for attribute selection are sparsity and algorithmic stability. An attribute selection algorithm is called sparse if it finds the sparsest or nearly sparsest set of attributes subject

to performance constraints (e.g., small generalization error) (Vaiciukynas *et al.,* 2014). An algorithm is called stable if it produces similar outputs when fed with similar inputs—that is, it selects similar attribute sets for two similar datasets (Xu *et al.,* 2012). As noted in Vaiciukynas *et al.* (2014), Xu *et al.* (2012), stability and sparsity constitute a trade-off. An algorithm that is sparse may be incapable of selecting similar sets of attributes across runs (Vaiciukynas *et al.,* 2014).

EAs for attribute selection vary on the number of objectives to be optimized, integration with other stages, and distribution of base learners. In Peimankar *et al.* (2016, 2017) a multi-objective Particle Swarm Optimization algorithm provided different attribute subsets to heterogeneous base learners, in order to predict whether power transforms will fail in the near future. In Sikdar *et al.* (2016, 2014b, 2015), two Pareto-based multi-objective differential evolution algorithms performs attribute selection, and then linear voting weight optimization, in a pipeline fashion (the generation stage is performed before the integration stage).

The encoding used in Kim *et al.* (2002) considers each individual as an ensemble of classifiers. Classifiers are trained differently based on their input features. Each classifier competes with its neighbors within the same ensemble; and at a higher level, ensembles compete among themselves based on their predictive accuracy.

4.3 Model optimization

Models may have their hyper-parameters and/or structure modified while creating a pool or ensemble of base learners. We divide this category of our taxonomy into two groups: pre-model and post-model optimization.

Pre-model optimization involves fine-tuning the hyper-parameters of the base learners that will generate the base models. We call these approaches *pre-model* because the optimization happens prior to model generation. Examples are: tuning a neural network's learning rate; a SVM's type of kernel function (Rosales-Pérez *et al.,* 2017); L2 regularization (Woon & Kramer, 2016); and random forests' number of trees (Saha *et al.,* 2016).

Pre-model approaches may support heterogeneous sets of base learners. For example, in Saha *et al.* (2016) the authors select both the types of base learner and their respective hyper-parameters, together with a set of attributes that will be assigned to a given learner. They used the NSGA-II (Deb *et al.,* 2002) algorithm, and the one-point crossover keeps base models and hyper-parameters together, only allowing to swap the selected attributes for each model.

Post-model approaches try to improve an existing model. Examples are layout and inner node selection for decision trees (Augusto *et al.,* 2010; Wen & Ting, 2016; Mauša & Grbac, 2017), and topology, weight, and activation function optimization in neural networks (Fernández *et al.,* 2016b; Ojha *et al.,* 2017). Weights are also optimized in Krithikaa and Mallipeddi (2016), where an ensemble of heterogeneous parametric models are optimized by differential evolution.

Post-model encoding depends on the type of base learner being used, and so are more common on homogeneous sets of base learners. In Kim and Cho (2008b), the weights of ANNs are modified by a GA. The authors adopt a matrix of size $W \times W$, where W is the number of neurons in the entire network. The upper diagonal encodes whether two given neurons are connected, and the lower diagonal encodes the weights associated with those connections.

Some studies perform both pre- and post-model optimization. In Ojha *et al.* (2017), first the topology of a neural network is evolved by using NSGA-II. The best found topology then has its parameters (e.g., weights and activation functions) adjusted by a multi-objective Differential Evolution method. In the end, the final population is submitted to a voting scheme optimized by another EA.

Attribute selection is often coupled with model optimization. In Tian and Feng (2014), both post-model optimization of Radial Basis Function Neural Networks and attribute selection were used, by performing both approaches in two subpopulations of the Cooperative Coevolutionary EA. In Rapakoulia *et al.* (2014), solutions for both tasks were placed within the same chromosome: in a 132-wide chromosome array, 88 bits are designated for attribute selection, 10 bits represent parameter *nu*

Table 4. *Categorization of studies that employ EAs in the generation stage of supervised ensemble learning*

Method	Related work
Instance selection	Krawczyk *et al.* (2016), Krawczyk and Woźniak (2014), Almeida and Galvão (2016), Gu and Jin (2014), Karakatič *et al.* (2015), de Lima and Ludermir (2014), Galar *et al.* (2013), Vluymans *et al.* (2016), Adair *et al.* (2017)
Attribute selection	Krawczyk *et al.* (2015), Peimankar *et al.* (2017, 2016), Sikdar *et al.* (2016), Saha *et al.* (2016), Oehmcke *et al.* (2015), Chen and Zhao (2008), Gu and Jin (2014), Lima and Ludermir (2015), Sikdar *et al.* (2015), Winkler *et al.* (2015), Sikdar *et al.* (2014b, 2012, 2014a, 2013), Kim *et al.* (2002), Liu *et al.* (2007), Mehdiyev *et al.* (2015), Chyzhyk *et al.* (2015), Zagorecki (2014), Vaiciukynas *et al.* (2014), Tian and Feng (2014), Rapakoulia *et al.* (2014), Debie *et al.* (2013b, a), Kumar and Kumar (2013), Aliakbarian and Fanian (2013), Bagheri *et al.* (2013), Tan *et al.* (2014), Chen *et al.* (2014), de Lima *et al.* (2014), De Lima and Ludermir (2013), Batista *et al.* (2017), Das *et al.* (2017)
Pre-model optimization	Woon and Kramer (2016), Rosales-Pérez *et al.* (2017), Saha *et al.* (2016), Almeida and Galvão (2016), Ojha *et al.* (2017), Khamis *et al.* (2016), Lima and Ludermir (2015), Winkler *et al.* (2015), Connolly *et al.* (2011, 2012), Pagano *et al.* (2012), Connolly *et al.* (2013), Kapp *et al.* (2010, 2011), Rosales-Pérez *et al.* (2014), Vaiciukynas *et al.* (2014), Rapakoulia *et al.* (2014), Dehuri *et al.* (2013), Singh *et al.* (2016), de Lima *et al.* (2014), De Lima and Ludermir (2013), Liu *et al.* (2017), Liew *et al.* (2017), Batista *et al.* (2017)
Post-model optimization	Krithikaa and Mallipeddi (2016), Fernández *et al.* (2016b), Wen and Ting (2016), Mauša and Grbac (2017), Augusto *et al.* (2010), Ojha *et al.* (2017), Chen *et al.* (2007), Lacy *et al.* (2015b), Davidsen and Padmavathamma (2015), Lacy *et al.* (2015a), Lévesque *et al.* (2012), Kim and Cho (2008b), Folino *et al.* (2010, 2007a, b, 2006), Duell *et al.* (2006), De Stefano *et al.* (2014), Stefano *et al.* (2011), Castro and Von Zuben (2011), Escovedo *et al.* (2013a,c,b, 2014), Mabu *et al.* (2014, 2015), Roebber (2015), Garg and Lam (2015), Trivedi and Dey (2014), Lones *et al.* (2014), de Lima and Ludermir (2014), Ishibuchi and Yamamoto (2003), Cao *et al.* (2013b, a), Santu *et al.,* (2014), Tian and Feng (2014), Kiranyaz *et al.* (2014), Schuman *et al.* (2014), Bhowan *et al.* (2011, 2013), Veeramachaneni *et al.* (2013), Debie *et al.* (2013b, a), Kaiping *et al.* (2013), Redel-Macías *et al.* (2013), Fernández *et al.* (2016a), Singh *et al.* (2016), Dufourq and Pillay (2014), de Lima *et al.* (2014), De Lima and Ludermir (2013), Vukobratović and Struharik (2017)

and threshold (integer and decimal part), 14 bits correspond to the gamma value and 20 bits are used for the parameter C of a nu-SVR learner.

Table 4 summarizes the work on EAs for the generation step of supervised ensemble learning, based on the taxonomy proposed in this section.

5. The selection stage of ensemble learning

From the pool of generated base models, model selection (or model pruning Parhizkar & Abadi, 2015a) is performed in order to define the final set of base models for the ensemble. Selection may be regarded as a multi-objective problem, where two objectives—predictive performance, and diversity—must be optimized. When the size of an ensemble is large, selecting base models based on these metrics can be computationally expensive if all ensemble options are considered, thus making the use of meta-heuristics (such as evolutionary algorithms) attractive (Parhizkar & Abadi, 2015a). However, simpler options (such as simply selecting the Φ most accurate learners) can also be used. Selection is often viewed as an optional stage and frequently not performed by traditional methods (e.g., boosting Freund & Schapire, 1995, bagging Breiman, 1996) or EA-based ones (e.g., Cao *et al.,* 2013b; Zagorecki, 2014).

Whether or not to perform selection is an issue for debate, with some authors proposing to bypass this stage (i.e., using the entire pool of models as ensemble) (Trawiński *et al.,* 2014). Lacy *et al.* (2015b) argue that model selection is irrelevant for ensemble learning, and that it is sufficient to select the Φ best models from the pool. According to Lacy *et al.* (2015b), Freund and Schapire (1996), Cagnini *et al.* (2018), from a predictive performance standpoint, this approach would be more effective than building an ensemble while considering diversity metrics. Other authors claim that there is little correlation between ensemble diversity and accuracy (Opitz, 1999; Breiman, 2001; Parhizkar & Abadi, 2015a; Cagnini *et al.,* 2018). On the other hand, some authors argue the opposite: for example, for regression, Wang and Alhamdoosh (2013) argue that the Φ best neural networks may not produce an ensemble with better mean squared error (MSE). This is also stated by Liu *et al.* (2015), adding that simply selecting the most accurate models may result in loss of predictive performance given that most of those models may be strongly correlated, leaving the opinion of the minority of the committee underrepresented.

Although Lacy *et al.* (2015b) and Liu *et al.* (2015) have different opinions on the utility of model selection, both agree that diversity measures are not a good proxy for ensemble quality, with Liu *et al.* (2015) suggesting that accuracy on a validation[5] set is sufficient. The rationale for using diversity measures is that by sacrificing individual accuracy for group diversity, one can achieve better group accuracy (Castro & Von Zuben, 2011; Parhizkar & Abadi, 2015a). Diversity in this case should not be measured at the genotype level (e.g., individuals encoding different attributes for the same base model), but rather measured based on the predictive performance of the algorithms decoded from the individuals. Diversity metrics can be of two types: pairwise or group-wise (Hernández *et al.,* 2015). A pairwise diversity metric often outputs a matrix of values denoting how diverse one base model is from another. Then, algorithms may select models that are, for example, more diverse to the other already-selected models. By contrast, group-wise metrics validate how diverse a group of base models is, thus requiring a previous strategy for composing groups. A review of diversity measures is presented by Hernández *et al.* (2015).

The motivations for using EAs for model selection are as follows. First, finding the optimal model subset within a large set is unfeasible with exhaustive search (the search space size is $\approx 2^B$, where B is the number of base models). By contrast, EAs perform a robust, global search for the near-optimal set of base models (Parhizkar & Abadi, 2015a). There is evidence that smaller ensembles can indeed outperform larger ones (Trawiński *et al.,* 2013). However, in practice, the optimal ensemble size varies across types of ensembles (e.g., bagging vs. boosting), types of base learners (with different biases), and datasets (with different degrees of complexity). Hence, given the complexity of the problem of selecting the optimal model size, and the typically large size of the search space, it is justifiable to use a robust search method like EAs to try to solve this problem.

Model selection can be further divided into two categories: static and dynamic selection (de Lima & Ludermir, 2014; Jackowski *et al.,* 2014; Jackowski, 2015; Cruz *et al.,* 2018). In static selection, regions of competence are defined at training time and are never changed (Jackowski *et al.,* 2014; Jackowski, 2015; Cruz *et al.,* 2018). In dynamic selection the regions are defined during classification time, through

[5] In supervised learning it is common to divide a dataset into three disjoint sets: training, validation and test. The validation set is used to evaluate the quality of models *while* training and helps to prevent overfitting to the training data. The test set is used for the final model evaluation *after* training.

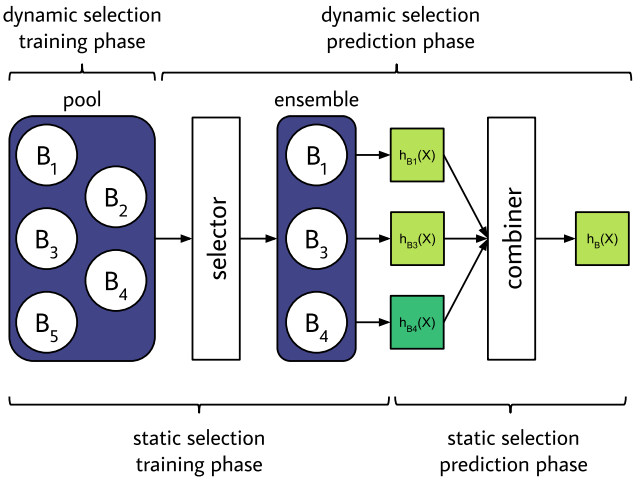

Figure 5. *Difference between static and dynamic selection strategies. While in static selection the competence estimator assigns regions to base learners during the training phase, in dynamic selection this is done during the prediction phase. Dynamic selection can also have a selector (e.g., oracle) that assigns a single base learner to regions of competence.*

the use of a competence estimator (Tsakonas & Gabrys, 2013; Jackowski *et al.,* 2014; Jackowski, 2015). Figure 5 puts both strategies in perspective.

Some studies in the literature (e.g., Britto *et al.,* 2014; Cruz *et al.,* 2018) experimentally assess whether dynamic selection methods are better than static selection ones. In Cruz *et al.* (2018), the authors compare static and dynamic selection methods on 30 different datasets, under the same protocol. The authors also compare these strategies with well-established ensemble algorithms, such as random forests, and AdaBoost. Only one of the 18 dynamic selection methods presented a worse predictive accuracy than simply using the best-performing classifier in the ensemble, and 66% of them outperformed a genetic algorithm performing static selection with majority voting as integration strategy. Furthermore, 44% and 61% of them presented a better average ranking than Random Forests and AdaBoost, respectively.

These results are also supported by Britto *et al.* (2014). Dynamic selection methods were statistically better than three other strategies: using the single best classifier in an ensemble, using all the generated classifiers, and using static selection methods. For the latter, dynamic selection algorithms won in 68% of the cases.

Note that some studies say they perform dynamic selection (e.g., Almeida & Galvão, 2016) via *k*-means, but in fact they perform static selection, since the assignment of classifiers is done during training time and does not change after that.

5.1 Static selection

Static selection is well-suited for batch-based learning, where the data distribution is not expected to change with time. Static selection assigns regions of competence during training time, thus allowing some freedom regarding which methods can be used. *Overproduce-and-select* is a traditional strategy for ensemble learning (Kapp *et al.,* 2010; Cordón & Trawiński, 2013), where an algorithm first generates a large pool of base models using a generation method (see Section 4). Then, the base models are selected from this pool and their votes are combined via an integration scheme (see Section 6). The rationale is that some models may perform poorly or have strongly correlated predictions, making some of these

safe for exclusion from the final ensemble (Trawiński *et al.,* 2014). EAs for this strategy aim at selecting the set of models that optimize a given criterion(a), often used as the fitness function.

A second strategy for static selection, known as *clustering-and-selection*, uses a clustering algorithm to assign models to distinct regions of competence in the training phase. In the testing phase, a new instance is submitted to the base model that covers the region closer to that instance. Studies using this strategy include (Rahman & Verma, 2013b, 2013a; e Silva *et al.,* 2013). In Jackowski *et al.* (2014), a GA was used for selecting the number of partitions in which the input space is divided. It then assigns an ensemble of classifiers to each partition, optimizing the voting weight of each base learner.

Overproduce-and-select and *clustering-and-select* differ regarding the region of competence where they will be employed. In the former, all selected base models will cast their predictions over the same region, whereas in the second they will be assigned to distinct ones.

In Wang and Alhamdoosh (2013) a hill-climbing strategy was used for increasing the size of the ensemble. By starting with only two classifiers (Neural Networks), the number of ensemble members is increased by adding classifiers that reduce the overall ensemble's error rate. In Dos Santos *et al.* (2008b), the authors investigate the impact of combining error rate (effectiveness), ensemble size (efficiency), and 12 diversity measures on the quality of static selection by using pairs of objectives. The authors also study conflicts between objectives, such as error rate/diversity measures and ensemble size/diversity measures. They argue that, among diversity measures, difficulty, inter-rate agreement, correlation coefficient, and double-fault are the best for combining with error rate, ultimately producing the best ensembles.

Studies that use the overproduce-and-select strategy often encode individuals as binary strings, where 0 denotes the absence of that model in the final ensemble and 1 the presence (Chen & Zhao, 2008). However, in Pourtaheri and Zahiri (2016) the individual size was doubled by using two values for each model: one for the aforementioned task, and another to determine the strength of that model's output in the final ensemble's prediction.

In Kim and Cho (2008a), attribute and model selection were performed at the same time. The authors use a binary matrix chromosome where each row represents a different base learner and each column a filter-based attribute selection approach. In this sense, if a bit is active somewhere within the individual's genotype, it means that the base learner of the corresponding row will be trained with the attributes selected by the filter approach of the corresponding column.

5.2 Dynamic selection

In dynamic selection, a single model or a subset of most competent learners is assigned to predict an unknown-class instance (Cruz *et al.,* 2018) (hereafter, unknown instance for short). This strategy is better suited for for example data stream learning, since the competence estimator naturally assigns base models to instances during the prediction phase. Dynamic selection was reported to perform better than boosting and static selection strategies (Cruz *et al.,* 2018). However, work on dynamic selection is much less frequent than work on static selection. Dynamic selection is also more computationally expensive, since estimators are required to define regions of competence for all predictions, which can be unfeasible in some cases (de Lima *et al.,* 2014; Britto *et al.,* 2014).

One approach for dynamic selection is to use random oracles (Cordón & Trawiński, 2013; Trawiński *et al.,* 2013). A random oracle is a mini-ensemble with only two base learners that are randomly assigned to competence regions (Trawiński *et al.,* 2013). At prediction time, the oracle decides which base learner to use for providing predictions for unknown instances.

Another strategy is to train a meta-learner. In Lima and Ludermir (2015), generation strategies of feature selection and pre-model optimization were combined with an overproduce-and-select strategy for generating a diverse pool of base learners. Next, a meta-learner was trained for selecting the best subset of models for predicting the class of unseen instances.

Table 5 shows the distribution of the surveyed EAs into the static and dynamic selection categories. The interested reader is referred to the work of Cruz *et al.* (2018) for a review on dynamic selection strategies.

Table 5. *Categorization of EAs for the selection stage of supervised ensemble learning*

Method	Related work
Static selection	Almeida and Galvão (2016), Milliken *et al.* (2016), Sikdar *et al.* (2016), Saha *et al.* (2016), Basto-Fernandes *et al.* (2016), Pourtaheri and Zahiri (2016), Obo *et al.* (2016), Oehmcke *et al.* (2015), Chen and Zhao (2008), Parhizkar and Abadi (2015a,b), Kim and Cho (2015), Jackowski *et al.* (2014), Jackowski (2015), Ko *et al.* (2006), Dos Santos *et al.* (2008a,b), Kim and Cho (2008a, 2005), Park and Cho (2003b,a), Coelho *et al.* (2003), Rosales-Pérez *et al.* (2014), Hernández *et al.* (2015), Chen and Yao (2006), Castro and Von Zuben (2011), Ma *et al.* (2015), Cordón and Trawiński (2013), Kumar and Kumar (2013), Wang and Alhamdoosh (2013), Tang *et al.* (2013), Singh *et al.* (2016), Chiu and Verma (2014), Dufourq and Pillay (2014), De Stefano *et al.* (2013), Rahman and Verma (2013b,a), e Silva *et al.* (2013), Shunmugapriya and Kanmani (2013), Liew *et al.* (2017), Asafuddoula *et al.* (2017), Batista *et al.* (2017), Basto-Fernandes *et al.* (2018)
Dynamic selection	Lima and Ludermir (2015), Cordón and Trawiński (2013), Trawiński *et al.* (2013)

6. The integration stage of ensemble learning

The last step of ensemble learning concerns the integration of votes (for classification) or value approximation (for regression) in order to maximize predictive performance. Ensemble integration, also called learner fusion (Trawiński *et al.,* 2014) or post-gate stage (Debie *et al.,* 2016), is the final chance to fine-tune the ensemble members in order to correct minor faults, such as giving more importance to a minority of learners that are however making correct predictions. Integration is an active research area in ensemble learning (Trawiński *et al.,* 2014). Similarly to the selection stage, this is another stage where using a validation set can be useful, since reusing the training set that was employed to generate base models can lead to overfitting.

As with other ensemble stages, there are two approaches for the integration of base learners: using traditional, non-EA methods, or using evolutionary algorithms. For classification, the most popular method is weighted majority voting, which allows to weight the contribution of each individual classifier to the prediction according to its competence via voting weights (Trawiński *et al.,* 2014):

$$h_B(X^{(i)}) = \underset{j}{\operatorname{argmax}} \left(\sum_{b=1}^{B} w_{b,j} \times [h_b(X) = c_j] \right) \tag{1}$$

where B is the number of classifiers, $w_{b,j}$ the weight associated with the bth classifier for the jth class, and $[h_b(X) = c_j]$ outputs 1 or 0 depending on the result of the Boolean test. This strategy has been shown to perform better than majority voting and averaging (Lacy *et al.,* 2015b). A simplification of that function sets all weights to 1, which turns this method into a simple majority voting, another popular approach (Zhang *et al.,* 2014). For instance, bagging uses a simple majority voting scheme, whereas boosting uses weighted majority voting (Zhang *et al.,* 2016b).

For regression, the most popular is the simple mean rule, which averages the predictions of base regressors, $h_B(X^{(i)}) = \frac{1}{B} \sum_{b=1}^{B} h_b(X^{(i)})$, where B is the number of regressors and $h_b(X^{(i)})$ is the prediction for the bth regressor. Simple aggregation strategies are better suited for problems where all predictions have comparable performance, however those methods are very vulnerable to outliers and unevenly performing models (Ma *et al.,* 2015). Other traditional methods for integrating regressors are average, weighted average, maximum, minimum, sum, and product rules (Kittler *et al.,* 1998; Mehdiyev *et al.,* 2015; Lacy *et al.,* 2015a).

However, there are plenty of studies that employ EAs for integrating base learners. These studies can be divided into two categories: optimizing the voting weights of a weighted majority voting rule; or optimizing/selecting the meta-models that will combine the outputs of base learners. Both categories may be interpreted as using meta-models for this task, as in stacking (Tsakonas & Gabrys, 2013; Mehdiyev *et al.,* 2015). Ensembles that use stacking are referred to as two-tier (or two-level) ensembles (Tsakonas & Gabrys, 2013). Those ensembles are well-suited, for example, for incremental learning (Tsakonas & Gabrys, 2013). Actually, when updating an existing ensemble model to consider new data, we may need to train only a few novel base models covering the new data and then re-train the meta-model with the both the novel and the previous base models. This is more efficient than re-training all existing base models in a single-level ensemble (Tsakonas & Gabrys, 2013). Two-tier ensembles were reported to perform better than simple weighting strategies in larger datasets (Neoh *et al.,* 2015). As disadvantages, two-tier ensembles are more susceptible to overfitting when compared to traditional integration methods and also increase the training time of the entire ensemble (Ma *et al.,* 2015). In practice, whether stacking or traditional aggregation methods are better is heavily influenced by the input data (Neoh *et al.,* 2015). The following sections will review the available methods that use EAs for the integration of base learners.

6.1 Linear models

Evolutionary algorithms in this category are concerned in learning a set of voting weights that will be used in a weighted majority voting integration strategy. A wide variety of methods were proposed for this task, such as using genetic algorithms (Krawczyk *et al.,* 2016; Ojha *et al.,* 2017), particle swarm optimization (Saleh *et al.,* 2016), flower pollination (Zhang *et al.,* 2017a), differential evolution (Sikdar *et al.,* 2012; Zhang *et al.,* 2016b, 2017b), etc. Those methods can be applied to both homogeneous (Chaurasiya *et al.,* 2016; Zhang *et al.,* 2016b, 2017b) and heterogeneous (Zhang *et al.,* 2014; Kim & Cho, 2015; Ojha *et al.,* 2015) base learner sets. For classification, methods may also differ in the number of voting weights, either by using one voting weight *per* classifier (e.g., Zhang *et al.,* 2014; Obo *et al.,* 2016) or one voting weight *per* classifier *per* class (e.g., Fatima *et al.,* 2013; Sikdar *et al.,* 2015; Davidsen & Padmavathamma, 2015).

For a thorough experimental analysis of both linear and nonlinear voting schemes, the reader is referred to the work of Lacy *et al.* (2015a), which presents the most comprehensive experimental comparison of EA-based combining methods to date. Notwithstanding, in the next sections we present a broader review of EAs proposed for this task, as well as methods that were not presented in Lacy *et al.* (2015a).

6.2 Nonlinear models

Instead of optimizing weights, one can use nonlinear models for integrating predictions. As the name implies, a nonlinear integration model does not use a set of voting weights (one for each model) to cast predictions, but instead relies on another arrangement to combine votes. As it is shown in this section, the types of nonlinear integration models used in literature may range from neural networks, to expression trees. Nonetheless, these nonlinear integration models may better exploit classifiers' diversity and accuracy properties (Escalante *et al.,* 2013).

6.2.1 Expression trees

One of the most popular EA-based nonlinear methods are expression trees (Escalante *et al.,* 2013; Tsakonas, 2014; Liu *et al.,* 2014a, 2015; Lacy *et al.,* 2015b,a; Ali & Majid, 2015; Folino *et al.,* 2016). Expression trees have models in their leaves and combination operators in their inner nodes.

For the problem of microarray data classification, in Liu *et al.* (2015, 2014a) some decision trees (initially trained with bagging) are fed to a Genetic Programming algorithm, which then induces a

population of expression trees (each allowed to have at most 3 levels) for combining the base classifiers' votes. After the evolutionary process is completed, expression trees with accuracy higher than the average are selected by a forward-search algorithm to compose the final meta-committee, which will predict the class of unknown instances.

6.2.2 Genetic fuzzy systems

Genetic fuzzy systems are popular in ensemble learning, where fuzzy systems optimized by EAs are used to predict the class of unknown instances. A study reports that fuzzy combiners can outperform crisp combiners in several scenarios (Trawiński *et al.,* 2014). There are several steps in the induction of fuzzy systems where EAs may be used: from tuning fuzzy membership functions to inducing rule bases (Cordón & Trawiński, 2013; Tsakonas & Gabrys, 2013). For instance, in Cordón and Trawiński (2013), Trawiński *et al.* (2014), a GA was used with a sparse matrix for codifying features and linguistic terms; and in Tsakonas and Gabrys (2013) a GP algorithm was used to evolve combination structures of a fuzzy system.

6.2.3 Neural networks

In an empirical work comparing several integration methods (Lacy *et al.,* 2015a), a multilayer percetron was used as a combination strategy. The output from base classifiers was used as input for the neural network, with an EA used for optimizing the weights of connections between neurons.

6.2.4 Evolutionary algorithms for selecting meta-combiners

In Shunmugapriya and Kanmani (2013), besides using the Artificial Bee Colony (ABC) algorithm for selecting base classifiers, the authors also use another ABC for selecting the meta-learner that will combine the votes of ensemble members.

6.3 Other methods

6.3.1 Induced Ordered Weighted Averaging (IOWA)

Ordered Weighted Averaging (OWA) (Yager, 1988) is a family of operators designed to combine several criteria in a multi-criteria problem. Let $A_1, A_2, A_3, \ldots, A_z$ be z criteria to be fulfilled in a multi-criteria decision function, and let A_j be how much a given solution fulfills the j-th criterion, $A_j \in [0, 1], \forall j = 1, \ldots, z$. The problem is then how to measure and compare solutions. This is solved by employing the OWA operators. OWA will combine two sets of values, a set of weights $W_1, W_2, \ldots, W_z, W_j \in (0, 1), \forall j = 1, \ldots, z, \sum_{j=1}^{z} W_j = 1$, and the set of ordered criteria $B = decreasing\_sort(A)$, by using a dot product, $F(A) = \sum_{j=1}^{z} W_j B_j$, with $F(A)$ as the fulfillment score of the solution. OWA is deemed ordered because weights are associated with the position in the combination function, rather than a specific criterion. For performing the combination, the criteria A are ordered based on their fulfillment rate (that is, the criterion that was most satisfied is combined with the first weight; the second most fulfilled criterion is combined with the second weight; and so on).

A method called Induced Ordered Weighted Averaging, or IOWA, is concerned with inducing the set of weights W, based on observational data (e.g., a dataset). In Bazi *et al.* (2014) a Multi-Objective EA based on Decomposition (MOEA-D) is used for inducing these weights, and IOWA is used to combine predictions from a set of Gaussian Process Regressors (GPR).

6.3.2 Error Correcting Output Codes (ECOC)

Error Correcting Output Codes (ECOC) (Bautista *et al.,* 2011) is a meta-method which combines many binary classifiers in order to solve multi-class problems (Bagheri *et al.,* 2013). It is an alternative to other multi-class strategies for binary classifiers (Cao *et al.,* 2014)—such as one-vs-one, which learns

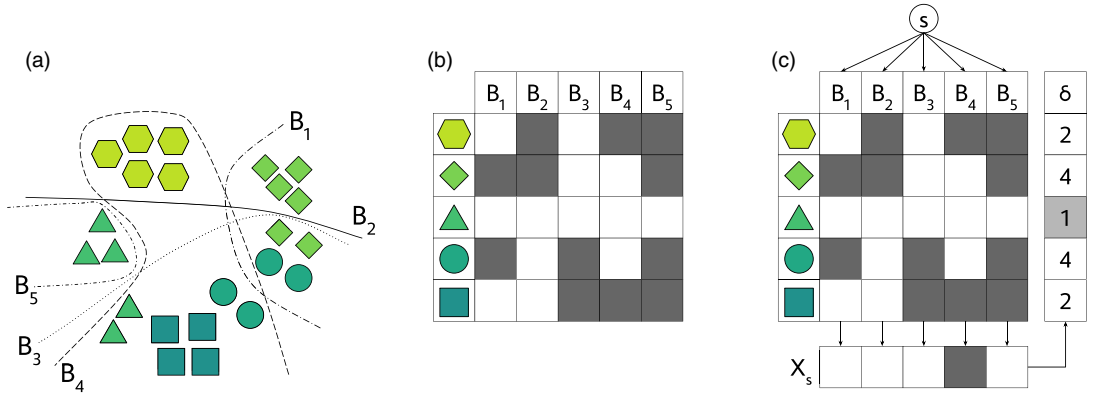

Figure 6. *(a) Feature space and decision boundaries of base classifiers. (b) Coding matrix, where black and white cells correspond to positive and negative classes, respectively, denoting the two partitions to be learned by each base classifier. (c) Decoding step, where the classifiers' predictions* $\{b_1, b_2, \ldots, b_5\}$ *for a given sample* s *are compared to the codewords* $\{y_1, \ldots, y_N\}$ *and* s *is labeled as the class codeword at minimum distance. Adapted from Bautista et al.* (2011).

a classifier for each pair of classes; and one-vs-all, which learns one classifier *per* class, discriminating instances from that class (positives) from all other instances (negatives). ECOC provides meta-classes to its classifiers (i.e., positive and negative classes are in fact combinations of instances from one or more classes). An example of ECOC is shown in Figure 6.

ECOC comprises two steps: encoding and decoding. The aim of encoding is to design a discrete decomposition matrix (codematrix) for the given problem (Bagheri *et al.,* 2013). A study reports that larger matrices (with regard to number of classifiers) improve predictive performance (Bagheri *et al.,* 2013). In the decoding phase, each classifier casts a vote to a meta-class for an unknown instance. The predicted class is computed by comparing the distance of the outputed codeword for that instance with the codeword from each real class via a similarity metric.

Though in classification we wish to reach top predictive accuracy, other measures should also be considered for evaluating the ECOC matrix, such as row separation and column diversity (Cao *et al.,* 2014). By using an indicator-based selection EA (IBEA), in Cao *et al.* (2014) the ensemble accuracy, individual classifier accuracy, and hamming distance were used as objectives for optimizing the layout of ECOC matrices, by manipulating the distribution of classes among base classifiers. In Bagheri *et al.* (2013), on the other hand, an attribute selection strategy was used to generate classifiers to be integrated by an ECOC scheme; hence, this work is labeled as a generation technique instead of an integration one.

To summarize, Table 6 shows the categorization of studies using EAs in the integration stage of ensemble learning, based on the type of integration method.

Sections 4, 5, and the current Section 6 have discussed EAs for each of the three stages of ensemble learning. In the next two sections, we focus on broader aspects of EAs for ensemble learning that are not specific to any single stage.

7. Objective (or fitness) functions

The fitness function is an essential component of an EA, since it defines the objective(s) to be optimized and guides the search accordingly. Due to the complexity of ensemble learning, there are several types of objectives that can be optimized. In this section we first review separately each of four broad types of objective (fitness) functions: effectiveness, efficiency, diversity, and complexity. Next, we review multi-objective optimization approaches.

Table 6. *Categorization of studies using EAs in the integration stage of ensemble learning*

Method	Related work
First degree polynomial	Zhang *et al*. (2016a), Haque *et al*. (2016), Zhang *et al*. (2016b, 2017b), Krawczyk *et al*. (2016, 2015), Krawczyk and Woźniak (2014), Krawczyk *et al*. (2013), Krawczyk and Schaefer (2014), Krawczyk *et al*. (2014), Chaurasiya *et al*. (2016), Obo *et al*. (2016), Zhang *et al*. (2017a), Onan *et al*. (2016), Pourtaheri and Zahiri (2016), Saleh *et al*. (2016), Basto-Fernandes *et al*. (2016), Ojha *et al*. (2017), Davidsen and Padmavathamma (2015), Lacy *et al*. (2015a), Kim and Cho (2015), Sikdar *et al*. (2015, 2014b), Jackowski *et al*. (2014), Jackowski (2014, 2015), Ojha *et al*. (2014, 2015), Schaefer (2013), Wozniak (2009), Sikdar *et al*. (2012, 2014a, 2013), Kim and Cho (2008a), Liu *et al*. (2007), Escovedo *et al*. (2013a, 2014, 2013c,b), Fuqiang *et al*. (2014), Zhang *et al*. (2014), Liu *et al*. (2014b), Fatima *et al*. (2013), Joardar *et al*. (2017), Galar *et al*. (2013), Cagnini *et al*. (2018)
Expression trees	Folino *et al.* (2016), Lacy *et al.* (2015b,a), Ali and Majid (2015), Liu *et al.* (2015, 2014a), Tsakonas (2014), Escalante *et al.* (2013)
Genetic Fuzzy System	Cordón and Trawiński (2013), Trawiński *et al.* (2014), Tsakonas and Gabrys (2013)
Error-Correcting Output Codes	Cao *et al.* (2014)
Artificial Neural Network	Lacy *et al.* (2015a)
IOWA	Bazi *et al.* (2014)
Meta-learner selection	Shunmugapriya and Kanmani (2013)

7.1 Effectiveness, diversity, complexity, and efficiency

An objective function measures *effectiveness* when it evaluates the ensemble's predictive accuracy. This is essential to ensemble learning and is addressed by all surveyed studies. The most popular objectives within this category are accuracy (or its dual, error rate) for classification tasks and MSE for regression tasks. The well-known accuracy measure is given by:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively. The error rate is simply: $1 - accuracy$. Note that accuracy and error rate have the drawback of not being suitable for highly imbalanced class distributions (Lévesque *et al.,* 2012), since they are relatively easy to optimize by predicting nearly always the majority class.

The MSE is given by:

$$\text{MSE}(X^{(i)}) = \frac{1}{N}\sum_{i=1}^{N}(h_B(X^{(i)}) - Y^{(i)})^2 \tag{3}$$

which computes the difference between the predicted value $h_B(X^{(i)})$ and the real value $Y^{(i)}$, for all N instances. Other effectiveness measures include exponential squared loss (Joardar *et al.,* 2017), geometric mean (Cao *et al.,* 2013a,b; Rapakoulia *et al.,* 2014; Vluymans *et al.,* 2016), imbalance ratio

(Vluymans *et al.,* 2016), and confidence (Kim & Cho, 2008a, 2015; Singh *et al.,* 2016), to name just a few. In general, such measures have the advantage of coping better with imbalanced class distributions than the aforementioned accuracy measure (or its dual error rate).

A diversity metric evaluates how diverse an ensemble's members (base learners) are. A diversity measure is often used as an objective in the selection stage of ensemble learning (Section 5), and it can also used in the generation stage (Section 4). We refer the reader to Castro and Von Zuben (2011) for a review on diversity measures for generating models; and next we discuss the controversial issue of using diversity as an objective in general, regardless of the ensemble learning stage.

Several researchers defend diversity as a valid objective (e.g., Folino *et al.,* 2006; Chen & Zhao, 2008; Stefano *et al.,* 2011; Chiu & Verma, 2013), stating that it contributes to ensemble accuracy (Chiu & Verma, 2013). De Stefano *et al.* (2011) state that, as the number of base learners increase, so does the probability that a minority of correct base learners will be overrun by a majority of wrong base learners, and thus the need for using diversity measures to reverse that effect. Also, in EAs, genetic material from well-performing solutions tend to be propagated to their offspring, often compromising diversity (Duell *et al.,* 2006).

However, other researchers do not see the utility of diversity measures (e.g., Ko *et al.,* 2006; Lacy *et al.,* 2015b; Liu *et al.,* 2015), stating that the correlation between ensemble accuracy and diversity is not as strong as expected (Trawiński *et al.,* 2013). Some authors also note that classic ensemble learning methods (e.g., bagging, boosting, and random subspace) introduce diversity in an ensemble without directly measuring it (Dos Santos *et al.,* 2008b). We can conclude, from this debate, that the relationship between ensemble effectiveness and diversity is not fully understood yet (Dos Santos *et al.,* 2008b, Trawiński *et al.,* 2013).

Diversity can be measured based on the ensemble's characteristics encoded in an individual's genotype, or based on the predictions made by each base learner. In the latter case, a set of base learners is said to be diverse when their errors are not correlated (Saleh *et al.,* 2016). Examples of diversity measures in this category include Yule's Q statistic (Tian & Feng, 2014), average residual correlation coefficient (Tsakonas, 2014), and negative correlation (Bhowan *et al.,* 2011, 2013). The most popular measure seems to be the Kohavi-Wolpert variance (Rahman & Verma, 2013b, 2013a; Chiu & Verma, 2014), given by:

$$KW = \frac{1}{NB^2} \sum_{j=1}^{N} L(X^{(j)})(B - L(X^{(j)})) \tag{4}$$

where B is the number of classifiers, N is the number of instances in the (training or validation) evaluation set, and $L(X^{(j)})$ is the number of classifiers within the ensemble that correctly predict the class of instance $X^{(j)}$.

Diversity measures can be divided into pairwise and group measures. The latter evaluate diversity among all classifiers in the ensemble, whereas pairwise metrics evaluate diversity between two classifiers and require an averaging technique for obtaining a group measure from all classifier-pairwise measures (Hernández *et al.,* 2015). This is performed by the disagreement measure. The pairwise disagreement measure (Castro & Von Zuben, 2011) is defined as:

$$\text{Diff}(B_i, B_j) = \frac{L^{01} + L^{10}}{L^{00} + L^{01}L^{10} + L^{11}} \tag{5}$$

where B_i and B_j are, respectively, the i-th and j-th classifiers within the ensemble, L^{10} is the number of instances correctly classified by B_i and wrongly classified by B_j, and so on for the remaining indices L^{01}, L^{00}, L^{11}. Pairwise disagreement varies from 0 to 1, with 0 indicating no disagreement (i.e., equal predictions) and 1 maximum disagreement. The plain disagreement measure (Liu *et al.,* 2007) simply averages the overall disagreement among the members of the ensemble:

$$\text{PSM} = \sum_{i=1}^{B} \sum_{j=i+1}^{B} \sum_{k=1}^{N} \frac{\text{Diff}(B_i, B_j)}{((B-1) \times B \times N)} \tag{6}$$

where B is the number of classifiers, N the number of instances in the training or validation set. For an extensive list of diversity measures for ensemble learning, the reader is referred to Kuncheva and Whitaker (2003), Ko *et al.* (2006), Hernández *et al.* (2015).

Complexity metrics evaluate how complex the classifiers in the ensemble, or the ensemble as a whole, are. The most popular complexity metrics are the number of activated classifiers (for classifier selection) (Park & Cho, 2003b; Ishibuchi & Yamamoto, 2003; Kim & Cho, 2005, 2008a; Dos Santos *et al.,* 2008b; Trawiński *et al.,* 2013; Cordón & Trawiński, 2013; Trawiński *et al.,* 2014) and the number of attributes used by the models induced by the base learners (Chen & Yao, 2006; Aliakbarian & Fanian, 2013; Tan *et al.,* 2014; Chen *et al.,* 2014; Zagorecki 2014; Rapakoulia *et al.,* 2014; Lima & Ludermir, 2015; Sikdar *et al.,* 2015; Winkler *et al.,* 2015). Other complexity metrics include the number of nodes in flexible neural trees (Ojha *et al.,* 2017); the number of hidden neurons in a neural network (Connolly *et al.,* 2013; Lima & Ludermir, 2015); the structured minimization principle (Garg & Lam, 2015); the number of support vectors in a SVM model (Rapakoulia *et al.,* 2014); and the length of fuzzy rules (Ishibuchi & Yamamoto, 2003). Most of these are measures of the size of an ensemble or its base members, so they are simple to compute; but the trade-off is that they may not capture a more sophisticated aspect of complexity (like complex interactions between the ensemble's base members).

Efficiency is a desired objective when an ensemble must be fast, during training and/or testing (prediction) phase. Training efficiency is obviously important in very large datasets. In addition, both training and testing efficiency are especially important in data stream scenarios, where a continuous flow of incoming data is presented to the system, and predictions must be made in a real-time basis. As a fitness function, training and test time also have the advantage of being very easy to compute; but they can introduce a trade-off between computational time and effectiveness, which could reduce the EA's effectiveness.

Complexity and efficiency metrics are related since, broadly speaking, reducing the complexity of the base learners or the ensemble as a whole leads to more efficient ensemble learning systems—for example, reducing the number of base learners (a complexity metric) leads to faster ensembles, for a fixed type of base learner. Note, however, that the number of base learners is not directly a measure of efficiency, since efficiency depends on both the number and the type of base learners. For example, an ensemble with a given number N of decision tree algorithms would probably be trained faster than an ensemble with $N/2$ neural networks, since the latter type of base learner is much slower than the former. In addition, it is possible to improve efficiency without directly reducing the complexity of the models in the ensemble—for example, by reducing the number of instances fed to the ensemble learning system.

In our survey, only two studies optimize efficiency, one measuring prediction time reduction (Oehmcke *et al.,* 2015), and the other measuring training set size reduction (for instance selection) (Rosales-Pérez *et al.,* 2017), as a proxy for training time. None of the surveyed EAs employed training time *per se* as a metric.

7.2 Single vs. multi-objective optimization

Ensemble learning methods performing single-objective optimization are obviously constrained to optimize effectiveness. However, ensemble learning may be naturally viewed as a multi-objective task, involving also other types of objectives. Figure 7 shows the distribution of other objectives that were optimized along effectiveness in studies that employed multiple objectives.

In this work we follow the taxonomy of multi-objective optimization approaches proposed in Freitas (2004), where approaches are categorized into three types: (i) weighted fitness functions, where each objective is assigned a user-defined (typically, very ad hoc) weight indicating that objective's importance; (ii) the lexicographic approach, where the user only ranks the objectives in terms of their priorities (no ad hoc numerical weights), and then the EA selects individuals for reproduction by trying to optimize the objectives in their decreasing order of priority; and (iii) the Pareto dominance approach, where the EA evolves a set of non-dominated solutions in the Pareto sense—that is, a solution is non-dominated if it is not worse than any other according to each objective *and* it is better than others according to at least one objective.

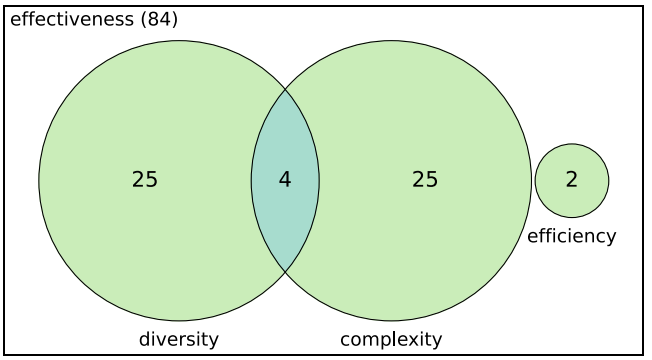

Figure 7. *Distribution of objectives across EAs using multiple objectives. Effectiveness (predictive performance) is optimized in all* 84 *studies. From these,* 56 *work optimize another objective, be it diversity, complexity, or efficiency. Only four studies (Dos Santos et al. 2008b; Cordón & Trawiński, 2013; Trawiński et al. 2013; Ojha et al. 2017) optimize three objectives (effectiveness, diversity, and complexity), and no study optimizes all four at the same time.*

In the surveyed papers, the least popular approach was the lexicographic one (Lima & Ludermir, 2015), followed by weighted fitness functions (Park & Cho, 2003b; Coelho *et al.,* 2003; Kim & Cho, 2005, 2008a, 2015; Parhizkar & Abadi, 2015a, 2015b; Hernández *et al.,* 2015; Winkler *et al.,* 2015; Zhang *et al.,* 2016a; Liew *et al.,* 2017; Batista *et al.,* 2017), then the single-objective approach (see the single-objective entry in Table 7) and finally the Pareto dominance approach as the most popular one (all the papers that were not cited in this paragraph and are within the multi-objective entry in Table 7).

8. Types of evolutionary algorithms

A variety of EAs have been employed in ensemble learning. While some of these EAs are less frequent in literature (e.g., Flower Pollination Algorithm Zhang *et al.,* 2017a, Levy-Flight Firefly Algorithm Zhang *et al.,* 2016a), others are more common. Among them, Genetic Algorithms (GAs) is the most popular, followed by Genetic Programming, and Differential Evolution.

Within GAs, apart from its *vanilla* version, Non-dominated Sorting Genetic Algorithm II (NSGA-II) is the most popular implementation. This choice seems due to NSGA-II's ability to deal with multiple objectives, suiting well the multi-objective nature of ensemble learning. Table 8 shows the distribution of the surveyed studies according to the type of EA used.

9. Types of base learners

In the surveyed studies, the most commonly used type of base learner is ANNs, used in 75 studies; followed by tree-based algorithms (e.g., decision trees, arithmetic trees), used in 48 studies; and SVMs, used in 46 studies. The number of studies using each type of base learner algorithm is shown in Table 9.

Some ensemble techniques are more appropriate for some type(s) of selected base learner(s). SVMs, for instance, are stable classifiers, making the techniques of selecting either instances or attributes for each base learner inefficient as a diversity inductor (Batista *et al.,* 2017).

The majority of the studies use homogeneous ensembles, as opposed to heterogeneous ones (113 vs. 47). Four work use both types of ensembles, which brings the count to 117 and 51, respectively. Figure 8 shows the base learners that were used in at least 5 studies, as well as the study's ensemble type: either homogeneous, or heterogeneous.

Table 7. *Studies categorized by number and type of objectives employed*

Number of objectives	Nature	Related work
Single-objective	Effectiveness	Chaurasiya *et al.* (2016), Krithikaa and Mallipeddi (2016), Zhang *et al.* (2017a), Haque *et al.* (2016), Almeida and Galvão (2016), Folino *et al.* (2016), Khamis *et al.* (2016), Chen and Zhao (2008), Davidsen and Padmavathamma (2015), Lacy *et al.* (2015a), Kim and Cho (2015), Karakatič *et al.* (2015), Ali and Majid (2015), Liu *et al.* (2015, 2014a), Krawczyk *et al.* (2013), Krawczyk and Schaefer (2014), Krawczyk *et al.* (2014), Jackowski *et al.* (2014), Jackowski (2015), Ojha *et al.* (2014, 2015), Schaefer (2013), Wozniak (2009), Dos Santos *et al.* (2008a), Connolly *et al.* (2011, 2012), Pagano *et al.* (2012), Kapp *et al.* (2010, 2011), Sikdar *et al.* (2012, 2013), Kim *et al.* (2002), Park and Cho (2003a), Kim and Cho (2008b), Folino *et al.* (2010, 2007a,b, 2006), Liu *et al.* (2007), Duell *et al.* (2006), De Stefano *et al.* (2014), Stefano *et al.* (2011), Escovedo *et al.* (2013a,c,b, 2014), Mabu *et al.* (2014, 2015), Roebber (2015), Ma *et al.* (2015), Chyzhyk *et al.* (2015), Trivedi and Dey (2014), Fuqiang *et al.* (2014), Zhang *et al.* (2014), Lones *et al.* (2014), Cao *et al.* (2013b,a), Vaiciukynas *et al.* (2014), Kiranyaz *et al.* (2014), Schuman *et al.* (2014), Shunmugapriya and Kanmani (2013), Fatima *et al.* (2013), Joardar *et al.* (2017), Dehuri *et al.* (2013), Tsakonas and Gabrys (2013), Cordón and Trawiński (2013), Kaiping *et al.* (2013), Redel-Macías *et al.* (2013), Bagheri *et al.* (2013), e Silva *et al.* (2013), Tang *et al.* (2013), Fernández *et al.* (2016a), Singh *et al.* (2016), Dufourq and Pillay (2014), Vukobratović and Struharik (2017), Liu *et al.* (2017), Batista *et al.* (2017), Cagnini *et al.* (2018), Chen *et al.* (2007), (Liu *et al.* 2014b), Veeramachaneni *et al.* (2013), De Stefano *et al.* (2013), Woon and Kramer (2016)
Multi-objective	Effectiveness	Wen and Ting (2016), Zhang *et al.* (2016b, 2017b), Obo *et al.* (2016), Sikdar *et al.* (2016), Saha *et al.* (2016), Fernández *et al.* (2016b), Zhang *et al.* (2016a), Rosales-Pérez *et al.* (2017), Mauša and Grbac (2017), Augusto *et al.* (2010), Ojha *et al.* (2017), Peimankar *et al.* (2017, 2016), Pourtaheri and Zahiri (2016), Milliken *et al.* (2016), Saleh *et al.* (2016), Onan *et al.* (2016), Basto-Fernandes *et al.* (2016), Krawczyk *et al.* (2016, 2015), Krawczyk and Woźiak (2014), Oehmcke *et al.* (2015), Gu and Jin (2014), Lacy *et al.* (2015b), Lima and Ludermir (2015), Parhizkar and Abadi (2015a,b), Kim and Cho (2015), Sikdar *et al.* (2015), Winkler *et al.* (2015), Jackowski (2014), Ko *et al.* (2006), Dos Santos *et al.* (2008b), Connolly *et al.* (2013), Lévesque *et al.* (2012), Sikdar *et al.* (2014a), Kim and Cho (2008a), Park and Cho (2003b), Coelho *et al.* (2003), Rosales-Pérez *et al.* (2014), Hernández *et al.* (2015), Chen and Yao (2006), Liu *et al.* (2007), Castro and Von Zuben (2011), Mehdiyev *et al.* (2015),

Table 7. *Continued*

Number of objectives	Nature	Related work
		Cao *et al*. (2014), Garg and Lam (2015), Trawiński *et al*. (2014), de Lima and Ludermir (2014), Ishibuchi and Yamamoto (2003), Zagorecki (2014), Santu *et al*. (2014), Tian and Feng (2014), Bazi *et al*. (2014), Tsakonas (2014), Rapakoulia *et al*. (2014), Escalante *et al*. (2013), Rahman and Verma (2013b,a), Bhowan *et al*. (2011), Cordón and Trawiński (2013), Debie *et al*. (2013b,a), Kumar and Kumar (2013), Aliakbarian and Fanian (2013), Wang and Alhamdoosh (2013), Galar *et al*. (2013), Trawiński *et al*. (2013), Vluymans *et al*. (2016), Chiu and Verma (2014), Tan *et al*. (2014), Chen *et al*. (2014), de Lima *et al*. (2014, De Lima and Ludermir (2013), Liew *et al*. (2017), Asafuddoula *et al*. (2017), Batista *et al*. (2017), Adair *et al*. (2017), Basto-Fernandes *et al*. (2018), Das *et al*. (2017), Sikdar *et al*. (2014b), Bhowan *et al*. (2013), Kim and Cho (2005)
	Efficiency	Rosales-Pérez *et al.* (2017), Oehmcke *et al.* (2015)
	Diversity	Peimankar *et al.,* (2017, 2016), Krawczyk *et al.,* (2015, 2014, 2016), Ojha *et al.* (2017), Gu and Jin (2014), Lacy *et al.* (2015b), Parhizkar and Abadi (2015a,b), Ko *et al.* (2006), Dos Santos *et al.* (2008b), Coelho *et al.* (2003), Hernández *et al.* (2015), Liu *et al.* (2007), Castro and Von Zuben (2011), Tian and Feng (2014), Tsakonas (2014), Rahman and Verma (2013b, a), Bhowan *et al.* (2011), Cordón and Trawiński (2013), Wang and Alhamdoosh (2013), Galar *et al.* (2013), Trawiński *et al.* (2013), Chiu and Verma (2014), Batista *et al.* (2017), Bhowan *et al.* (2013), Kim and Cho (2005)
	Complexity	Saha *et al.* (2016), Zhang *et al.* (2016a), Ojha *et al.* (2017), Lima and Ludermir (2015), Sikdar *et al.* (2015), Winkler *et al.* (2015), Connolly *et al.* (2013), Rosales-Pérez *et al.* (2014), Chen and Yao (2006), Garg and Lam (2015), Ishibuchi and Yamamoto (2003), Zagorecki (2014), Rapakoulia *et al.* (2014), Aliakbarian and Fanian (2013), Tan *et al.* (2014), Chen *et al.* (2014), Liew *et al.* (2017), Basto-Fernandes *et al.* (2018), Obo *et al.* (2016), Pourtaheri and Zahiri (2016), Kim and Cho (2015), Dos Santos *et al.* (2008b), Kim and Cho (2008a), Park and Cho (2003b), Trawiński *et al.* (2014), Cordón and Trawiński (2013), Trawiński *et al.* (2013), Santu *et al.* (2014), Asafuddoula *et al.* (2017)

Homogeneous ensembles are composed by the same base learner paradigm. However, this is not to say that all base learners are exactly the same. When using neural networks, those models can have distinct activation functions or topologies. According to Rahman and Verma (2013c), there are five strategies for inducing diverse models in homogeneous ensembles: (i) post-model optimization (see Section 4.3); (ii) manipulation of the error function; (iii) distinct attribute subsets across base learners (see Section

Table 8. *Studies organized by the type of EA employed, and the ensemble learning stage optimized*

Evolutionary family	Generation	Selection	Integration
Flower pollination			Zhang *et al.* (2017a)
Clonal selection	Batista *et al.* (2017)	Batista *et al.* (2017)	
Evolutionary algorithm	Redel-Macías *et al.* (2013)		
Inclined planes optimization		Pourtaheri and Zahiri (2016)	Pourtaheri and Zahiri (2016)
Multi-objective EA		Basto-Fernandes *et al.* (2018, 2016)	Basto-Fernandes *et al.* (2016)
Moth-flame optimization			Zhang *et al.* (2016a)
Levy-flight firefly algorithm			Zhang *et al.* (2016a)
Virus-evolutionary genetic algorithm			Fuqiang *et al.* (2014)
Many-objectives evolutionary algorithm		Asafuddoula *et al.* (2017)	
Evolutionary strategy	Vukobratović and Struharik (2017), Woon and Kramer (2016)		
Artificial bee colony	Cao *et al.* (2013b,a)	Parhizkar and Abadi (2015b), Shunmugapriya and Kanmani (2013), Parhizkar and Abadi (2015a)	Joardar *et al.* (2017), Shunmugapriya and Kanmani (2013)
Estimation of distribution algorithm	Escovedo *et al.* (2013c), Chen and Zhao (2008), Escovedo *et al.* (2013b), Castro and Von Zuben (2011), Escovedo *et al.* (2013a 2014)	Chen and Zhao (2008), Castro and Von Zuben (2011)	Escovedo *et al.* (2013c), Cagnini *et al.* (2018), Escovedo *et al.* (2013b,a, 2014)
Particle swarm optimization	Chen *et al.* (2007), Peimankar *et al.* (2017), Connolly *et al.* (2011), Chen and Zhao (2008), Kiranyaz *et al.* (2014), Peimankar *et al.* (2016), Connolly *et al.* (2012), Kapp *et al.* (2011), Pagano *et al.* (2012), Connolly *et al.* (2013), (Kapp *et al.* (2010), Batista *et al.* (2017)	Dos Santos *et al.* (2008a), Chen and Zhao (2008), Pourtaheri and Zahiri (2016), Batista *et al.* (2017)	Saleh *et al.* (2016), Pourtaheri and Zahiri (2016), Ali and Majid (2015)

Table 8. *Continued*

Evolutionary family	Generation	Selection	Integration
Differential evolution	Vluymans *et al.* (2016), Sikdar *et al.* (2013), de Lima and Ludermir (2014), de Lima *et al.* (2014), Sikdar *et al.* (2012), Debie *et al.* (2013b), Vaiciukynas *et al.* (2014), Dehuri *et al.* (2013), Sikdar *et al.* (2016), De Lima and Ludermir (2013), Krithikaa and Mallipeddi (2016), Sikdar *et al.* (2015, 2014a,b), Lima and Ludermir (2015)	De Stefano *et al.* (2013), Sikdar *et al.* (2016), Lima and Ludermir (2015)	Sikdar *et al.* (2013), Zhang *et al.* (2016b, 2017b), Onan *et al.* (2016), Zhang *et al.* (2014), Haque *et al.* (2016), Sikdar *et al.* (2012, 2015, 2014a,b), Chaurasiya *et al.* (2016)
Genetic programming	Bhowan *et al.* (2011), Lacy *et al.* (2015a), Folino *et al.* (2007a), Bhowan *et al.* (2013), Chen *et al.* (2007), Roebber (2015), Dufourq and Pillay (2014), Garg and Lam (2015), Wen and Ting (2016), Lones *et al.* (2014), Stefano *et al.* (2011), Mabu *et al.* (2014), Folino *et al.* (2010), Mabu *et al.* (2015), Folino *et al.* (2007b,2006), Trivedi and Dey (2014), Veeramachaneni *et al.* (2013), Lacy *et al.* (2015b), De Stefano *et al.* (2014)	Dufourq and Pillay (2014)	Lacy *et al.* (2015a), Liu *et al.* (2014a), Tsakonas (2014, 2013), Liu *et al.* (2015), Escalante *et al.* (2013), Ali and Majid (2015), Lacy *et al.* (2015b), Folino *et al.* (2016)
Genetic algorithms	Almeida and Galvão (2016), Chen *et al.* (2014), Kim and Cho (2008b), Lacy *et al.* (2015a), Fernández *et al.* (2016a), Bagheri *et al.* (2013), Kumar and Kumar (2013), Dufourq and Pillay (2014), Mehdiyev *et al.* (2015),	Kim and Cho (2005), Hernández *et al.* (2015), Almeida and Galvão (2016), Dos Santos *et al.* (2008a), (Obo *et al.* 2016), Kumar and Kumar (2013), Dufourq and Pillay (2014),	Krawczyk *et al.* (2013), Krawczyk and Schaefer (2014), Lacy *et al.* (2015a), Trawiński *et al.* (2014), Cao *et al.* (2014), Obo *et al.* (2016), Kim and Cho (2015), Ojha *et al.* (2017),

Table 8. *Continued*

Evolutionary family	Generation	Selection	Integration
	Mauša and Grbac (2017), Tian and Feng (2014), Saha *et al.* (2016), Augusto *et al.* (2010), Zagorecki (2014), Karakatič *et al.* (2015), Aliakbarian and Fanian (2013), Tan *et al.* (2014), Ojha *et al.* (2017), Liew *et al.* (2017), Krawczyk *et al.* (2016), Krawczyk *et al.* (2015), Kaiping *et al.* (2013), Liu *et al.* (2017), Santu *et al.* (2014), Rapakoulia *et al.* (2014), Gu and Jin (2014), Duell *et al.* (2006), Krawczyk and Woźniak (2014), Debie *et al.* (2013b), Schuman *et al.* (2014), Trivedi and Dey (2014), Kim *et al.* (2002), Rosales-Pérez *et al.* (2017), Winkler *et al.* (2015), Rosales-Pérez *et al.* (2014), Khamis *et al.* (2016), Galar *et al.* (2013), Lévesque *et al.* (2012), Oehmcke *et al.* (2015), Chyzhyk *et al.* (2015), Fernández *et al.* (2016b), Das *et al.* (2017), Vluymans *et al.* (2016), Liu *et al.* (2007), Ishibuchi and Yamamoto (2003), Davidsen and Padmavathamma (2015), Adair *et al.* (2017), Batista *et al.* (2017)	e Silva *et al.* (2013), De Stefano *et al.* (2013), Saha *et al.* (2016), Park and Cho (2003a), Chiu and Verma (2014), Kim and Cho (2015), Ma *et al.* (2015), Liew *et al.* (2017), Rahman and Verma (2013b), Basto-Fernandes *et al.* (2018), Jackowski *et al.* (2014), Kim and Cho (2008a), Wang and Alhamdoosh (2013), Jackowski (2015), Tang *et al.* (2013), Dos Santos *et al.* (2008b), Rosales-Pérez *et al.* (2014), Chen and Yao (2006), Trawiński *et al.* (2013), Oehmcke *et al.* (2015), Ko *et al.* (2006), Cordón and Trawiński (2013), Rahman and Verma (2013a), Coelho *et al.* (2003), Basto-Fernandes *et al.* (2016), Park and Cho (2003b), Milliken *et al.* (2016), Batista *et al.* (2017)	Krawczyk *et al.* (2016, 2015), Jackowski *et al.* (2014, 2014), Schaefer (2013), Kim and Cho (2008a), Bazi *et al.* (2014), Jackowski (2015), Krawczyk and Woźniak (2014), Fatima *et al.* (2013), Liu *et al.* (2014b), Galar *et al.* (2013), Wozniak (2009), Krawczyk *et al.* (2014), Ojha *et al.* (2014), Ojha *et al.* (2015), Cordón and Trawiński (2013), Liu *et al.* (2007), Davidsen and Padmavathamma (2015), Basto-Fernandes *et al.* (2016)

Table 9. *Studies organized according to the base learners they employ. We only show in this table base learners that are present in at least 5 papers. For the complete list of base learners, please refer to our website at https://henryzord.github.io/eael*

Base learner	Related work
Gaussian process regression	Winkler *et al.* (2015), Ojha *et al.* (2015), Ojha *et al.* (2014), Bazi *et al.* (2014), Liu *et al.* (2017)
Linear regression	Jackowski (2015), Ojha *et al.* (2015), Jackowski *et al.* (2014), Ojha *et al.* (2014), Lévesque *et al.* (2012)
Fuzzy rule-based classifier	Trawiński *et al.* (2014), Cordón and Trawiński (2013), Mehdiyev *et al.* (2015), Ishibuchi and Yamamoto (2003), Fernández *et al.* (2016a), Trawiński *et al.* (2013)
Conditional random fields	Sikdar *et al.* (2013), Sikdar *et al.* (2012), Fatima *et al.* (2013), Sikdar *et al.* (2015), Sikdar *et al.* (2014b), Sikdar *et al.* (2014a), Sikdar *et al.* (2016)
Logistic regression	Onan *et al.* (2016), Escalante *et al.* (2013), Folino *et al.* (2016), Haque *et al.* (2016), Saha *et al.* (2016), Hernández *et al.* (2015), Shunmugapriya and Kanmani (2013)
Random forest	(Saha *et al.,* 2016, Milliken *et al.,* 2016, Ali and Majid 2015, Rosales-Pérez *et al.,* 2014, Escalante *et al.,* 2013, Winkler *et al.,* 2015, Vaiciukynas *et al.,* 2014, Hernández *et al.,* 2015)
Rule-based	Basto-Fernandes *et al.* (2016), Sikdar *et al.* (2016), Mabu *et al.* (2015), Mabu *et al.* (2014), Santu *et al.* (2014), Davidsen and Padmavathamma (2015), Debie *et al.* (2013a,b), Mehdiyev *et al.* (2015), Roebber (2015), Basto-Fernandes *et al.* (2018)
Naïve Bayes	Milliken *et al.* (2016), Ali and Majid (2015), Kumar and Kumar (2013), Haque *et al.* (2016), Folino *et al.* (2016), Zhang *et al.* (2014), Onan *et al.* (2016), Escalante *et al.* (2013), Zagorecki (2014), Peimankar *et al.* (2017), Peimankar *et al.* (2016), Shunmugapriya and Kanmani (2013), Jackowski (2014), Karakatič *et al.* (2015), Hernández *et al.* (2015), Das *et al.* (2017)
K-nearest neighbor	Shunmugapriya and Kanmani (2013), Folino *et al.* (2016), Escalante *et al.* (2013), Pourtaheri and Zahiri (2016), Peimankar *et al.* (2017), Peimankar *et al.* (2016), Krithikaa and Mallipeddi (2016), Oehmcke *et al.* (2015), Winkler *et al.* (2015), Ali and Majid (2015), Jackowski (2014), Jackowski *et al.* (2014), Dos Santos *et al.* (2008a,b), (Ko *et al.* (2006), Park and Cho (2003b,a), Vluymans *et al.* (2016), Kim and Cho (2015), Haque *et al.* (2016), Hernández *et al.* (2015), Saleh *et al.* (2016), Zhang *et al.* (2014), Liu *et al.* (2007), Kim and Cho (2005), Kim and Cho (2008a), Milliken *et al.* (2016), Asafuddoula *et al.* (2017), Das *et al.* (2017)
Support vector machines	Gu and Jin (2014), Saha *et al.* (2016), Peimankar *et al.* (2016), Rosales-Pérez *et al.* (2017), Krawczyk and Schaefer (2014), Parhizkar and Abadi (2015b), Kapp *et al.* (2010), Ali and Majid (2015), Rosales-Pérez *et al.* (2014), Saleh *et al.* (2016), Woon and Kramer (2016), Sikdar *et al.* (2012), Kim and Cho (2005), Asafuddoula *et al.* (2017), Kim and Cho (2015), Sikdar *et al.* (2013), Hernández *et al.* (2015), Batista *et al.* (2017), Tsakonas (2014), Sikdar *et al.* (2016), Kapp *et al.* (2011), Ojha *et al.* (2014), Park and Cho (2003b), Rahman and Verma (2013b,a), Escalante *et al.* (2013), Ojha *et al.* (2015), Rapakoulia *et al.* (2014), Park and Cho (2003a), Tsakonas and Gabrys (2013), Liu *et al.* (2017), Das *et al.* (2017), Vaiciukynas *et al.* (2014), Jackowski *et al.* (2014), Kim and Cho (2008a), Jackowski (2015), Zhang *et al.* (2016a), Haque *et al.* (2016), Winkler *et al.* (2015), Onan *et al.* (2016), Fatima *et al.* (2013), Cao *et al.* (2014), Peimankar *et al.* (2017), Chaurasiya *et al.* (2016), Coelho *et al.* (2003), Liu *et al.* (2007)

Table 9. *Continued*

Base learner	Related work
Trees	Karakatič *et al.* (2015), Saha *et al.* (2016), Augusto *et al.* (2010), Zhang *et al.* (2014), Bhowan *et al.* (2011), Trivedi and Dey (2014), Cagnini *et al.* (2018), Schaefer (2013), Bagheri *et al.* (2013), Rosales-Pérez *et al.* (2014), Milliken *et al.* (2016), Asafuddoula *et al.* (2017), Veeramachaneni *et al.* (2013), Folino *et al.* (2010), Krawczyk and Woźniak (2014), Krawczyk *et al.* (2015), Hernández *et al.* (2015), Folino *et al.* (2016), Tan *et al.* (2014), Folino *et al.* (2007a), De Stefano *et al.* (2014), Chen *et al.* (2014), Lones *et al.* (2014), Stefano *et al.* (2011), Krawczyk *et al.* (2016), Folino *et al.* (2006), Mauša and Grbac (2017), Das *et al.* (2017), Wen and Ting (2016), Krawczyk *et al.* (2014), Lacy *et al.* (2015b), Shunmugapriya and Kanmani (2013), Liu *et al.* (2015), Folino *et al.* (2007b), Haque *et al.* (2016), Garg and Lam (2015), Vukobratović and Struharik (2017), Winkler *et al.* (2015), Lévesque *et al.* (2012), Bhowan *et al.* (2013), Dufourq and Pillay (2014), Lacy *et al.* (2015a), Oehmcke *et al.* (2015), Aliakbarian and Fanian (2013), Galar *et al.* (2013), Krawczyk *et al.* (2013), Liu *et al.* (2014a, 2007)
Artificial neural network	Escovedo *et al.* (2014), Peimankar *et al.* (2016), Dehuri *et al.* (2013), Castro and Von Zuben (2011), Zhang *et al.* (2016b), Schuman *et al.* (2014), Chiu and Verma (2014), Bagheri *et al.* (2013), Rosales-Pérez *et al.* (2014), Tian and Feng (2014), Saleh *et al.* (2016), Almeida and Galvão (2016), Khamis *et al.* (2016), Kim *et al.* (2002), Wang and Alhamdoosh (2013), Fernández *et al.* (2016b), Connolly *et al.* (2013), Kim and Cho (2005), Asafuddoula *et al.* (2017), Kaiping *et al.* (2013), Kim and Cho (2015), Liew *et al.* (2017), De Stefano *et al.* (2013), Wozniak (2009), Hernández *et al.* (2015), Pourtaheri and Zahiri (2016), Chen *et al.* (2007), Ojha *et al.* (2017), Chen and Zhao (2008), Cao *et al.* (2013a), Connolly *et al.* (2012), Tsakonas (2014), Tan *et al.* (2014), Chen *et al.* (2014), Kiranyaz *et al.* (2014), Lones *et al.* (2014), Ojha *et al.* (2014), Park and Cho (2003b), De Lima and Ludermir (2013), Chyzhyk *et al.* (2015), Park and Cho (2003a), Ojha *et al.* (2015), Escalante *et al.* (2013), Chen and Yao (2006), Tsakonas and Gabrys (2013), Escovedo *et al.* (2013b), Obo *et al.* (2016), Liu *et al.* (2017), Das *et al.* (2017), Lacy *et al.* (2015b), Jackowski *et al.* (2014), Escovedo *et al.* (2013a), Kim and Cho (2008a), de Lima and Ludermir (2014), Kim and Cho (2008b), Jackowski (2015), Zhang *et al.* (2016a), Escovedo *et al.* (2013c), Haque *et al.* (2016), Zhang *et al.* (2017b), Winkler *et al.* (2015), Liu *et al.* (2014b), Lima and Ludermir (2015), Zhang *et al.* (2017a), e Silva *et al.* (2013), Fatima *et al.* (2013), de Lima *et al.* (2014), Tang *et al.* (2013), Cao *et al.* (2013b), Singh *et al.* (2016), Peimankar *et al.* (2017), Duell *et al.* (2006), Connolly *et al.* (2011), Pagano *et al.* (2012), Redel- Macías *et al.* (2013)

4.2); (iv) manipulation of output targets, in which some instances in the training set have their class labels switched, for inducing diversity; and (v) distinct instance subsets across base learners (see Section 4.1). Strategies (ii) and (iv) are not covered in this survey, though, due to the lack of relevant papers.

Heterogeneous ensembles comprise base learners from distinct paradigms. As such, there is no pressure for inducing diversity in the ensemble, since learners from distinct paradigms tend to make diverse predictions. In our survey, among the studies using diversity measures, 19 have a homogeneous set of base learners, while 9 use a heterogeneous set. These numbers include a single paper that proposes both homogeneous and heterogeneous ensembles. The larger number of papers with homogeneous sets is probably because it is simpler to work with homogeneous ensembles than with heterogeneous ones.

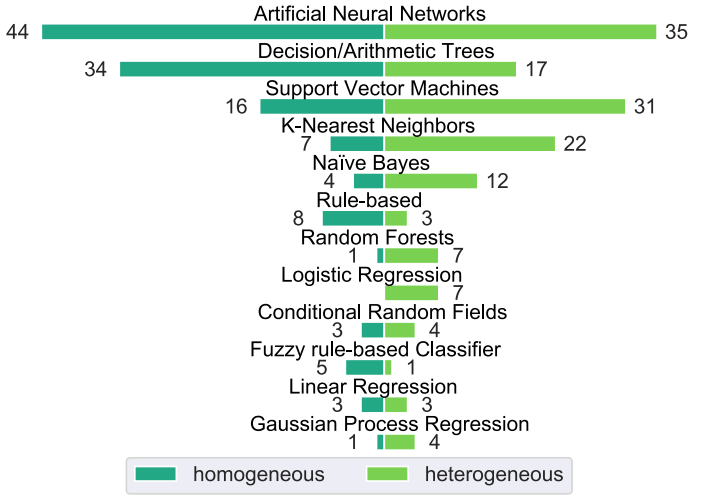

Figure 8. *Types of base learners used in surveyed work, as well as their distribution.*

While Figure 8 depicts an overview of the number of studies per type of base learners, Table 10 identifies which studies are using which types of base learners, and also their configuration (homogeneous or heterogeneous). Note that a few studies (4) use both types of ensemble.

10. Algorithm complexity by stages of ensemble learning

As it is expected with a survey that broadly reviews the literature, it is difficult to derive a single algorithm complexity for Ensemble Learning with Evolutionary Algorithms, or even multiple accurate estimations. For this reason, in this section we will discuss what aspects have the most impact on algorithm complexity, as well as deriving a general complexity for relevant ensemble stages (i.e., generation and selection).

The most time-consuming step in an evolutionary algorithm is evaluating candidate solutions. How candidate solutions are built differs among EAs, but it is safe to assume that these solutions are, in the context of ensemble learning, already-built, ready-to-deploy ensembles. Consider for example a 'generic' evolutionary algorithm for the generation stage. Let us assume that this EA does not have a pool of base classifiers—that is, for each candidate solution, it builds base classifiers that will only be used by that solution. If this EA runs for G generations and has a population of S individuals, then the time complexity of this algorithm is $O(G \times (BS\Psi\Omega))$, where B is the number of base classifiers in each ensemble (individual), Ψ is the complexity of the most time-consuming base classifier employed in the ensemble (in case it is heterogeneous), or the complexity of the only base classifier used (in case the ensemble is homogeneous), and Ω is the complexity of aggregating predictions among base classifiers. Ω can be as fast as $O(B)$, when using majority voting, to an arbitrarily complex algorithm, such as employing Genetic Algorithms to evolve Expression Trees (presented in Section 6.2.1). When more than one aggregation policy is available, the reader should assume the worst case scenario—that is, that all solutions will take as much time to train as it takes to use the most time-consuming base classifiers, coupled with the most time-consuming aggregation policy.

Note that having a pre-built pool of classifiers to choose from reduces the time complexity. In the case of a 'generic' EA for the selection stage, when the EA performs static selection, the time complexity is $O((P \times \Psi) + G \times (S\Omega))$, where P is the pool size, and $P \geq B$ (in this case, each individual will have a different B). Once base classifiers are built, their matrices of probabilities can be stored in memory and aggregated by any desired aggregation policy with Ω complexity. Since static selection consists only in flipping bits in a binary vector, its complexity is negligible. On the other hand, the complexity of

Table 10. *Studies organized according to the base learners' homogeneity/heterogeneity*

Homogeneity	Related work
Homogeneous	Mauša and Grbac (2017), Ojha *et al.* (2017), Joardar *et al.* (2017), Rosales-Pérez *et al.* (2017), Almeida and Galvão (2016), Wen and Ting (2016), Zhang *et al.* (2016b), Woon and Kramer (2016), Krawczyk *et al.* (2016), Fernández *et al.* (2016a), Zhang *et al.* (2017b), Fernández *et al.* (2016b), Vluymans *et al.* (2016), Obo *et al.* (2016), Chaurasiya *et al.* (2016), Singh *et al.* (2016), Khamis *et al.* (2016), Oehmcke *et al.* (2015), Gu and Jin (2014), Davidsen and Padmavathamma (2015), Lacy *et al.* (2015a), Parhizkar and Abadi (2015b), Sikdar *et al.* (2015), Karakatič *et al.* (2015), Liu *et al.* (2015), Krawczyk *et al.* (2015), Mabu *et al.* (2015), Ma *et al.* (2015), Roebber (2015), Cao *et al.* (2014), Chyzhyk *et al.* (2015), Garg and Lam (2015), Chiu and Verma (2014), Trivedi and Dey (2014), Jackowski *et al.* (2014), De Stefano *et al.* (2014), Fuqiang *et al.* (2014), Escovedo *et al.* (2014), Krawczyk *et al.* (2014), Jackowski (2014), Krawczyk and Schaefer (2014), Trawiński *et al.* (2014), Dufourq and Pillay (2014), de Lima and Ludermir (2014), de Lima *et al.* (2014), Liu *et al.* (2014a), Krawczyk and Woźniak (2014), Cao *et al.* (2013b), Zagorecki (2014), Mabu *et al.* (2014), Santu *et al.* (2014), Liu *et al.* (2014b), Vaiciukynas *et al.* (2014), Sikdar *et al.* (2014b,a), Tian and Feng (2014), Kiranyaz *et al.* (2014), Schuman *et al.* (2014), Rapakoulia *et al.* (2014), Bazi *et al.* (2014), Schaefer (2013), Krawczyk *et al.* (2013), Rahman and Verma (2013b), De Lima and Ludermir (2013), De Stefano *et al.* (2013), Rahman and Verma (2013a), Debie *et al.* (2013b), Debie *et al.* (2013a), Escovedo *et al.* (2013b), Galar *et al.* (2013), Trawiński *et al.* (2013), Escovedo *et al.* (2013c,a), Connolly *et al.* (2013), Cordón and Trawiński (2013), Wang and Alhamdoosh (2013), Aliakbarian and Fanian (2013), Cao *et al.* (2013a), Dehuri *et al.* (2013), Kumar and Kumar (2013), Kaiping *et al.* (2013), Veeramachaneni *et al.* (2013), Bhowan *et al.* (2013), Lévesque *et al.* (2012), Connolly *et al.* (2012), Pagano *et al.* (2012), Connolly *et al.* (2011), Stefano *et al.* (2011), Kapp *et al.* (2011), Bhowan *et al.* (2011), Kapp *et al.* (2010), Augusto *et al.* (2010), Folino *et al.* (2010), Wozniak (2009), Chen and Zhao (2008), Dos Santos *et al.* (2008a), Kim and Cho (2008b), Dos Santos *et al.* (2008b), Folino *et al.* (2007a), Chen *et al.* (2007), Folino *et al.* (2007b), Chen and Yao (2006), Duell *et al.* (2006), Ko *et al.* (2006), Folino *et al.* (2006), Ishibuchi and Yamamoto (2003), Kim *et al.* (2002), Vukobratović and Struharik (2017), Liew *et al.* (2017), Batista *et al.* (2017), Adair *et al.* (2017), Basto-Fernandes *et al.* (2018), Cagnini *et al.* (2018)
Heterogeneous	Peimankar *et al.* (2017), Zhang *et al.* (2017a), Haque *et al.* (2016), Pourtaheri and Zahiri (2016), Milliken *et al.* (2016), Saleh *et al.* (2016), Peimankar *et al.* (2016), Onan *et al.* (2016), Basto-Fernandes *et al.* (2016), Folino *et al.* (2016), Zhang *et al.* (2016a), Saha *et al.* (2016), Sikdar *et al.* (2016), Krithikaa and Mallipeddi (2016), Lima and Ludermir (2015), Parhizkar and Abadi (2015a), Hernández *et al.* (2015), Kim and Cho (2015), Winkler *et al.* (2015), Ali and Majid (2015), Mehdiyev *et al.* (2015), Ojha *et al.* (2015), Zhang *et al.* (2014), Lones *et al.* (2014), Ojha *et al.* (2014), Rosales-Pérez *et al.* (2014), Tsakonas (2014), Shunmugapriya and Kanmani (2013), Escalante *et al.* (2013), Fatima *et al.* (2013), e Silva *et al.* (2013), Bagheri *et al.* (2013), Sikdar *et al.* (2013), Tang *et al.* (2013), Redel- Macías *et al.* (2013), Tsakonas and Gabrys (2013), Sikdar *et al.* (2012), Castro and Von Zuben (2011), Kim and Cho (2008a), Liu *et al.* (2007), Kim and Cho (2005), Park and Cho (2003b,a), Coelho *et al.* (2003), Liu *et al.* (2017), Asafuddoula *et al.* (2017), Das *et al.* (2017)
Both	Lacy *et al.* (2015b), Jackowski (2015), Tan *et al.* (2014), Chen *et al.* (2014)

dynamic selection (presented in Section 5.2) lies on the complexity of training a selector to be later used during the prediction phase. In this case, the complexity of a generic EA performing dynamic selection is approximately $O((P \times \Psi) + G \times (S\Xi\Omega))$, where Ξ is the complexity of learning that selector.

11. Conclusions and new research directions

Ensemble learning is an extensive research field due to the improvement it presents in comparison to single learners and the easiness to integrate within some challenging types of machine learning problems (e.g., data stream learning and datasets with imbalanced class distributions Krawczyk *et al.,* 2016; Folino *et al.,* 2016; Mauša & Grbac, 2017). For some problems, inducing a single, stronger-than-all base learner can be a difficult task (Almeida & Galvão, 2016), while ensembles of models can perform better with regard to both effectiveness and efficiency (Breiman, 1996; Gu & Jin, 2014; Oehmcke *et al.,* 2015; Parhizkar & Abadi, 2015a; Krithikaa & Mallipeddi, 2016).

Ensemble learning can be further enhanced by using EAs in one or more of its learning stages: generation, selection, and integration. In this survey, we reviewed a large number of studies using many types of EAs for ensemble learning and proposed a taxonomy to classify such studies with regard to different aspects of ensemble learning. We also reviewed the debate on controversial topics, like the selection of ensemble members (rather than using all members) and the usefulness of optimizing a diversity measure for the members of the ensemble.

In order to facilitate the review of specific studies discussed in our survey, we make available the metadata used to compile our figures and tables. By using such metadata, one can see at a glance all the main aspects of a given study (e.g., which base learners, objectives, learning stages, etc, a study is using). The repository is available at https://github.com/henryzord/eael. We also provide a master table, in the form of a website, listing all surveyed work, and their classification according to our taxonomy: https://henryzord.github.io/eael.

11.1 Summary of findings

First, we discuss the main findings of this survey regarding each stage of the ensemble learning process. The generation stage, that is where the ensemble members are generated, was found to be the most popular step to employ EAs, having more studies dedicated to it than the selection and integration stages combined. Wrapper methods were found to be much more common than filter ones for the instance selection and attribute selection approaches. This seems natural, considering that, unlike filters, wrappers select attributes or instances customized for the supervised learning algorithm to be used later (to induce a model), which tends to improve predictive performance. However, wrappers are normally much slower than filters. Hence, in applications with large datasets or where efficiency is a critical factor, the filter approach deserves more attention. In addition, in the model tuning approach for generation, post-model optimization (used to improve an existing model) was found to be more popular than pre-model optimization (used before learning the model).

The next stage, selection—where ensemble members are selected to be used in the testing phase—is an optional stage, which is missing in many ensemble learning systems. In this stage, static selection, where the regions of competence of ensemble members are identified at training time, was found to be much more popular than dynamic selection, where those regions are identified at testing (prediction) time. This seems partly due to the greater simplicity and computational efficiency of the former, since dynamic selection in general requires a more time-consuming process of identifying regions of competence of ensemble members for each testing instance.

In the integration stage, where the predictions of the base learners are integrated into a final prediction for each instance, by far the most popular approach among the surveyed studies was the use of a first degree polynomial—a simple linear approach. Among the nonlinear integration techniques, the most common was the use of expression trees, using a genetic programming algorithm. It seems that more

research is needed on nonlinear techniques for integration, in order to determine whether or not their higher computational complexity could be justified by a significant increase in predictive performance.

Regarding the number of objectives in the fitness function, multi-objective EAs were found to be much more common than single-objective ones. This seems natural, given the multi-objective nature of the ensemble learning problem. In terms of specific types of objectives, effectiveness (predictive performance) is used by all surveyed EAs, since it is essential. Diversity and complexity share a second place, despite diversity being a controversial objective, as discussed earlier. Finally, efficiency, the capacity to generate ensembles that are computationally fast, is optimized in only two studies. Efficiency is important in huge datasets (since the training process must eventually finish, and a compromise between time spent in training stage and effectiveness must be made), and in data stream scenarios, where data is treated not as a fixed batch of instances, but instead as a continuous flow. Efficiency and effectiveness are competing objectives, since efficiency prioritizes models that are faster to train (and thus more likely to be simpler, less accurate models). However, if efficiency is a priority, one could use techniques such as parallelization of one of evolutionary algorithms' steps (Hauschild & Pelikan, 2011) to alleviate this competition.

Regarding the main types of EAs used in the surveyed studies, the most popular one was by far Genetic Algorithms (often NSGA-II, a multi-objective GA), followed by Genetic Programming and Differential Evolution.

Regarding the main type of base learner, the most popular one was ANNs, followed by decision trees and SVMs. The popularity of ANNs as base learners seems partly due to a long history of interaction in the EA and ANN research areas and partly due to the nature of ANNs, whose performance can often be improved when using ensembles. However, learning ensembles of ANNs or SVMs tends to be very computationally expensive. This problem is mitigated when learning an ensemble of decision trees (much faster base learners).

11.2 New research directions

One direction for future research is the automated selection of the best combination of ensemble algorithms and their hyper-parameter settings for a given input dataset. This is a complex optimization problem because, as discussed earlier, there are many types of ensembles (e.g., bagging, boosting, stacking, etc), and for each type of ensemble, many types of base (classification or regression) algorithms can be chosen. In addition, both the ensemble type and its base algorithm type(s) typically have many hyper-parameters, whose settings also have a large influence on the ensemble's predictive performance. All these choices of algorithms and hyper-parameter settings interact in a complex manner, and ideally all these choices should be made in a synergistic way, optimizing all these choices as a whole for the specific dataset provided as input by the user. Emerging research has addressed this complex optimization problem by doing a search in the space of different types of learning algorithms and their hyper-parameter settings, in order to automatically select the best combination of algorithm and hyper-parameter settings for an input dataset (Wistuba *et al.,* 2017; Kordík *et al.,* 2018). To the best of our knowledge, although there are several EA-based systems that address this problem by considering a search space with many types of supervised learning algorithms (e.g., Olson *et al.,* 2016; de Sá *et al.,* 2017), there are only two studies using an EA to address this problem by considering a search space focused on ensembles (Kordík *et al.,* 2018; Xavier-Júnior *et al.,* 2018). This seems an area with good potential for research growth.

Also, it is yet to be seen a framework that provides a synergistic integration of two or more ensemble learning stages (generation, selection, and integration). Even when a study addresses two or more stages, this is not done synergistically; base learners are first generated and later selected, or first selected and later integrated. We are not aware of any EA addressing all three stages.

Finally, ensemble learning with EAs would benefit from a unified, generic-purpose software tool, similarly to what WEKA (Witten *et al.,* 2016) and scikit-learn (Pedregosa *et al.,* 2011) do for machine learning in general and Tensorflow (Abadi *et al.,* 2015) for deep learning. This would greatly facilitate the task of comparing different strategies, for example, distinct approaches for generating or selecting

base learners while keeping the same fitness function. We believe the development of such a framework to be a major step forward to the evolutionary ensemble learning community.

11.3 Further readings

While this work is, up to our knowledge, the first to present a broad review of evolutionary algorithms for ensemble learning, the following literature can give a better understanding on topics related to ensemble learning, not necessarily involving evolutionary algorithms.

A closely related work to ours is the one of Yao and Islam (2008), which present a review of evolutionary algorithms for ensemble learning, although focusing only in work that uses ANNs as base classifiers.

For a general comprehension on ensemble learning, not necessarily involving evolutionary algorithms, Sagi and Rokach (2018) presents recent, state-of-the-art methods for ensemble learning. This is an update of another review on ensemble learning of the same author, presented in the work of Rokach (2010). While the former reviews generation and integration methods, as well as presenting the main challenges when building methods for ensemble learning, the later reviews integration and selection methods, and a discussion on ensemble diversity.

In the work of Oza and Tumer (2008) the authors review ensemble methods for solving real-world problems, such as remote sensing, person recognition, and medicine applications. Athar *et al.* (2017) review classifier ensembles for sentiment analysis. Data stream analysis with ensembles is reviewed both in the work of Gomes *et al.* (2017) and Krawczyk *et al.* (2017).

A broad review on classification ensembles and its applications is presented in Tabassum and Ahmed (2016), while a survey on regression ensembles is presented in Mendes-Moreira *et al.* (2012).

Regarding the generation stage, Olvera-Lopez *et al.* (2010) presents a review on instance selection methods, which can be used in this stage of ensemble learning; while Debie *et al.* (2016) review ensemble methods that focus on feature selection.

For the selection stage, Britto *et al.* (2014) presents a review on dynamic selection of classifiers, as well as a statistical comparison of results of the reviewed methods. An update of the reviewed methods, as well as the proposed taxonomy by the authors, is presented by Cruz *et al.* (2018).

Acknowledgments. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. The authors would also like to acknowledge FAPERGS for partially funding this research.

Conflicts of interest. The authors declare none.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng X. 2015. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/.

Adair, J., Brownlee, A., Daolio, F. & Ochoa, G. 2017. Evolving training sets for improved transfer learning in brain computer interfaces. In *International Workshop on Machine Learning, Optimization, and Big Data*, 186–197. Springer.

Albukhanajer, W. A., Jin, Y. & Briffa, J. A. 2017. Classifier ensembles for image identification using multi-objective pareto features. *Neurocomputing* **238**, 316–327.

Ali, S. & Majid, A. 2015. Can–Evo–Ens: classifier stacking based evolutionary ensemble system for prediction of human breast cancer using Amino Acid sequences. *Journal of Biomedical Informatics* **54**, 256–269.

Aliakbarian, M. S. & Fanian, A. 2013. Internet traffic classification usqing MOEA and online refinement in voting on ensemble methods. In *Iranian Conference on Electrical Engineering*, 1–6. IEEE.

Almeida, L. M. & Galvão, P. S. 2016. Ensembles with clustering-and-selection model using evolutionary algorithms. In *Brazilian Conference on Intelligent Systems*, 444–449. IEEE.

Asafuddoula, M., Verma, B. & Zhang, M. 2017. A divide-and-conquer based ensemble classifier learning by means of many-objective optimization. *IEEE Transactions on Evolutionary Computation* **22**(5), 762–777.

Athar, A., Butt, W. H., Anwar, M. W. & Latif, M. 2017. Exploring the ensemble of classifiers for sentimental analysis: A systematic literature review. In *International Conference on Machine Learning and Computing*, 410–414. ACM.

Augusto, D. A., Barbosa, H. J. C. & Ebecken, N. F. F. 2010. Coevolutionary multi-population genetic programming for data classification. In *Conference on Genetic and Evolutionary Computation*, 933–940. ACM.

Bagheri, M. A., Gao, Q. & Escalera, S. 2013. A genetic-based subspace analysis method for improving error-correcting output coding. *Pattern Recognition* **46**(10), 2830–2839.

Basto-Fernandes, V., Yevseyeva, I., Méndez, J. R., Zhao, J., Fdez-Riverola, F. & Emmerich, M. T. M. 2016. A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification. *Applied Soft Computing* **48**, 111–123.

Basto-Fernandes, V., Yevseyeva, I., Ruano-Ordás, D., Zhao, J., Fdez-Riverola, F., Méndez, J. R. & Emmerich, M. T. M. 2018. Quadcriteria optimization of binary classifiers: error rates, coverage, and complexity. In *EVOLVE – A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation VI*, Tantar, A.-A., Tantar, E., Emmerich, M., Legrand, P., Alboaie, L. & Luchian, H. (eds), 37–49. Springer.

Batista, J. d. O., Rodrigues, R. B. & Varejão, F. M. 2017. Soft computing classifier ensemble for fault diagnosis. In *International Symposium on Industrial Electronics*, 1348–1353. IEEE.

Bautista, M. Á., Pujol, O., Baró, X. & Escalera, S. 2011. Introducing the separability matrix for error correcting output codes coding. In *International Workshop on Multiple Classifier Systems*, 227–236. Springer.

Bazi, Y., Alajlan, N., Melgani, F., AlHichri, H. & Yager, R. R. 2014. Robust estimation of water chlorophyll concentrations with Gaussian process regression and IOWA aggregation operators. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(7), 3019–3028.

Bhowan, U., Johnston, M. & Zhang, M. 2011. Evolving ensembles in multi-objective genetic programming for classification with unbalanced data. In *Conference on Genetic and Evolutionary Computation*, 1331–1338. ACM.

Bhowan, U., Johnston, M. & Zhang, M. 2013. Comparing ensemble learning approaches in genetic programming for classification with unbalanced data. In *Conference on Genetic and Evolutionary Computation*, 135–136. ACM.

Breiman, L. 1996. Bagging predictors. *Machine Learning* **24**(2), 123–140.

Breiman, L. 2001. Random forests. *Machine Learning* **45**(1), 5–32.

Britto Jr., A. S., Sabourin, R. & Oliveira, L. E. S. 2014. Dynamic selection of classifiers – A comprehensive review. *Pattern Recognition* **47**(11), 3665–3680.

Cagnini, H., Basgalupp, M. & Barros, R. 2018. Increasing boosting effectiveness with estimation of distribution algorithms. In *Congress on Evolutionary Computation*, 1–8. IEEE.

Cao, J.-J., Kwong, S., Wang, R. & Li, K. 2014. An indicator-based selection multi-objective evolutionary algorithm with preference for multi-class ensemble. In *International Conference on Machine Learning and Cybernetics*, 147–152. IEEE.

Cao, P., Li, B., Zhao, D. & Zaiane, O. 2013a. A novel cost sensitive neural network ensemble for multiclass imbalance data learning. In *International Joint Conference on Neural Networks*, 1–8. IEEE.

Cao, P., Zhao, D. & Zaiane, O. 2013b. Measure optimized cost-sensitive neural network ensemble for multiclass imbalance data learning. In *International Conference on Hybrid Intelligent Systems*, 35–40. IEEE.

Castro, P. A. D. & Von Zuben, F. J. 2011. Learning ensembles of neural networks by means of a Bayesian artificial immune system. *IEEE Transactions on Neural Networks* **22**(2), 304–316.

Chaurasiya, R. K., Londhe, N. D. & Ghosh, S. 2016. Binary DE-based channel selection and weighted ensemble of SVM classification for novel brain–computer interface using Devanagari script-based P300 Speller Paradigm. *International Journal of Human–Computer Interaction* **32**(11), 861–877.

Chen, H. & Yao, X. 2006. Evolutionary multiobjective ensemble learning based on Bayesian feature selection. In *Congress on Evolutionary Computation*, 267–274. IEEE.

Chen, W.-C., Tseng, L.-Y. & Wu, C.-S. 2014. A unified evolutionary training scheme for single and ensemble of feedforward neural network. *Neurocomputing* **143**, 347–361.

Chen, Y., Yang, B. & Abraham, A. 2007. Flexible neural trees ensemble for stock index modeling. *Neurocomputing* **70**(4), 697–703.

Chen, Y. & Zhao, Y. 2008. A novel ensemble of classifiers for microarray data classification. *Applied Soft Computing* **8**(4), 1664–1669.

Chiu, C.-Y. & Verma, B. 2013. Effect of varying hidden neurons and data size on clusters, layers, diversity and accuracy in neural ensemble classifier. In *International Conference on Computational Science and Engineering*, 455–459. IEEE.

Chiu, C.-Y. & Verma, B. 2014. Multi-objective evolutionary algorithm based optimization of neural network ensemble classifier. In *International Conference on Signal Processing and Communication Systems*, 1–5. IEEE.

Chyzhyk, D., Savio, A. & Graña, M. 2015. Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM. *Neural Networks* **68**, 23–33.

Coelho, A. L. V., Lima, C. A. M. & Von Zuben, F. J. 2003. GA-based selection of components for heterogeneous ensembles of support vector machines. In *Congress on Evolutionary Computation*, 2238–2245. IEEE.

Connolly, J.-F., Granger, E. & Sabourin, R. 2011. Comparing dynamic PSO algorithms for adapting classifier ensembles in video-based face recognition. In *Workshop on Computational Intelligence in Biometrics and Identity Management*, 1–8. IEEE.

Connolly, J.-F., Granger, E. & Sabourin, R. 2012. Evolution of heterogeneous ensembles through dynamic particle swarm optimization for video-based face recognition. *Pattern Recognition* **45**(7), 2460–2477.

Connolly, J.-F., Granger, E. & Sabourin, R. 2013. Dynamic multi-objective evolution of classifier ensembles for video face recognition. *Applied Soft Computing* **13**(6), 3149–3166.

Cordón, O. & Trawiński, K. 2013. A novel framework to design fuzzy rule-based ensembles using diversity induction and evolutionary algorithms-based classifier selection and fusion. In *International Work-Conference on Artificial Neural Networks*, 36–58. Springer.

Cruz, R. M. O., Sabourin, R. & Cavalcanti, G. D. C. 2018. Dynamic classifier selection: recent advances and perspectives. *Information Fusion* **41**, 195–216.

Das, A. K., Das, S. & Ghosh, A. 2017. Ensemble feature selection using bi-objective genetic algorithm. *Knowledge-Based Systems* **123**, 116–127.

Davidsen, S. A. & Padmavathamma, M. 2015. Multi-modal evolutionary ensemble classification in medical diagnosis problems. In *International Conference on Advances in Computing, Communications and Informatics*, 1366–1370. IEEE.

De Lima, T. P. F. & Ludermir, T. B. 2013. Optimizing dynamic ensemble selection procedure by evolutionary extreme learning machines and a noise reduction filter. In *International Conference on Tools with Artificial Intelligence*, 546–552. IEEE.

de Lima, T. P. F. & Ludermir, T. B. 2014. Ensembles of evolutionary extreme learning machines through differential evolution and fitness sharing. In *International Joint Conference on Neural Networks*, 2677–2682. IEEE.

de Lima, T. P. F., Sergio, A. T. & Ludermir, T. B. 2014. Improving classifiers and regions of competence in dynamic ensemble selection. In *Brazilian Conference on Intelligent Systems*, 13–18. IEEE.

de Sá, A. G. C., Pinto, W. J. G. S., Oliveira, L. O. V. B. & Pappa, G. L. 2017. RECIPE: A grammar-based framework for automatically evolving classification pipelines. In *European Conference on Genetic Programming*, 246–261. Springer.

De Stefano, C., Cioppa, A. D. & Marcelli, A. 2013. Evolutionary approaches for pooling classifier ensembles: Performance evaluation. In *International Conference of Soft Computing and Pattern Recognition*, 309–314. IEEE.

De Stefano, C., Folino, G., Fontanella, F. & Di Freca, A. S. 2014. Using Bayesian networks for selecting classifiers in GP ensembles. *Information Sciences* **258**, 200–216.

Deb, K. 2011. Multi-objective optimisation using evolutionary algorithms: an introduction. In *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*, 3–34. Springer.

Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6** (2), 182–197.

Debie, E., Shafi, K., Lokan, C. & Merrick, K. 2013a. Performance analysis of rough set ensemble of learning classifier systems with differential evolution based rule discovery. *Evolutionary Intelligence* **6**(2), 109–126.

Debie, E., Shafi, K., Merrick, K. & Lokan, C. 2016. On taxonomy and evaluation of feature selection-based learning classifier system ensemble approaches for data mining problems. *Computational Intelligence* **33**(3), 554–578.

Debie, E. S., Shafi, K. & Lokan, C. 2013b. REUCS-CRG: reduct based ensemble of supervised classifier system with combinatorial rule generation for data mining. In *Conference on Genetic and Evolutionary Computation*, 1251–1258. ACM.

Dehuri, S., Jagadev, A. K. & Cho, S.-B. 2013. Epileptic seizure identification from electroencephalography signal using DE-RBFNs ensemble. *Procedia Computer Science* **23**, 84–95.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, 1–15. Springer.

Dos Santos, E. M., Oliveira, L. S., Sabourin, R. & Maupin, P. 2008a. Overfitting in the selection of classifier ensembles: a comparative study between PSO and GA. In *Conference on Genetic and Evolutionary Computation*, 1423–1424. ACM.

Dos Santos, E. M., Sabourin, R. & Maupin, P. 2008b. Pareto analysis for the selection of classifier ensembles. In *Conference on Genetic and Evolutionary Computation*, 681–688. ACM.

Duell, P., Fermin, I. & Yao, X. 2006. Speciation techniques in evolved ensembles with negative correlation learning. In *Congress on Evolutionary Computation*, 3317–3321. IEEE.

Dufourq, E. & Pillay, N. 2014. Hybridizing evolutionary algorithms for creating classifier ensembles. In *World Congress on Nature and Biologically Inspired Computing*, 84–90. IEEE.

e Silva, E. J. d. R., Ludermir, T. B. & Almeida, L. M. 2013. Clustering and selection using grouping genetic algorithms for blockmodeling to construct neural network ensembles. In *International Conference on Tools with Artificial Intelligence*, 420–425. IEEE.

Escalante, H. J., Acosta-Mendoza, N., Morales-Reyes, A. & Gago-Alonso, A. 2013. Genetic programming of heterogeneous ensembles for classification. In *Iberoamerican Congress on Pattern Recognition*, 9–16. Springer.

Escovedo, T., da Cruz, A., Vellasco, M. & Koshiyama, A. 2013a. NEVE: a neuro-evolutionary ensemble for adaptive learning. In *International Conference on Artificial Intelligence Applications and Innovations*, 636–645. Springer.

Escovedo, T., da Cruz, A. V. A., Vellasco, M. & Koshiyama, A. S. 2013b. Using ensembles for adaptive learning: a comparative approach. In *International Joint Conference on Neural Networks*, 1–7. IEEE.

Escovedo, T., da Cruz, A. V. A., Vellasco, M. M. & Koshiyama, A. S. 2013c. Learning under concept drift using a neuro-evolutionary ensemble. *International Journal of Computational Intelligence and Applications* **12**(4), 1340002.

Escovedo, T., da Cruz, A. A., Koshiyama, A., Melo, R. & Vellasco, M. 2014. NEVE++: a neuro-evolutionary unlimited ensemble for adaptive learning. In *International Joint Conference on Neural Networks*, 3331–3338. IEEE.

Fatima, I., Fahim, M., Lee, Y.-K. & Lee, S. 2013. Classifier ensemble optimization for human activity recognition in smart homes. In *International Conference on Ubiquitous Information Management and Communication*, 1–7. ACM.

Fernández, A., del Ro, S. & Herrera, F. 2016a. A first approach in evolutionary fuzzy systems based on the lateral tuning of the linguistic labels for big data classification. In *International Conference on Fuzzy Systems*, 1437–1444. IEEE.

Fernández, J. C., Cruz-Ramrez, M. & Hervás-Martnez, C. 2016b. Sensitivity versus accuracy in ensemble models of artificial neural networks from multi-objective evolutionary algorithms. *Neural Computing and Applications* **30**(1), 289–305.

Folino, G., Pisani, F. S. & Sabatino, P. 2016. An incremental ensemble evolved by using genetic programming to efficiently detect drifts in cyber security datasets. In *Conference on Genetic and Evolutionary Computation*, 1103–1110. ACM.

Folino, G., Pizzuti, C. & Spezzano, G. 2006. Improving cooperative GP ensemble with clustering and pruning for pattern classification. In *Conference on Genetic and Evolutionary Computation*, 791–798. ACM.

Folino, G., Pizzuti, C. & Spezzano, G. 2007a. An adaptive distributed ensemble approach to mine concept-drifting data streams. In *International Conference on Tools with Artificial Intelligence*, 183–188. IEEE.

Folino, G., Pizzuti, C. & Spezzano, G. 2007b. StreamGP: tracking evolving GP ensembles in distributed data streams using fractal dimension. In *Conference on Genetic and Evolutionary Computation*, 1751–1751. ACM.

Folino, G., Pizzuti, C. & Spezzano, G. 2010. An ensemble-based evolutionary framework for coping with distributed intrusion detection. *Genetic Programming and Evolvable Machines* **11** (2), 131–146.

Freitas, A. A. 2004. A critical review of multi-objective optimization in data mining: a position paper. *ACM SIGKDD Explorations Newsletter* **6**(2), 77–86.

Freund, Y. & Schapire, R. E. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, 23–37. Springer.

Freund, Y. & Schapire, R. E. 1996. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, 148–156. International Machine Learning Society.

Fuqiang, D., Minqing, Z. & Jia, L. 2014. Virus-evolutionary genetic algorithm based selective ensemble for steganalysis. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 553–558. IEEE.

Galar, M., Fernández, A., Barrenechea, E. & Herrera, F. 2013. EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition* **46**(12), 3460–3471.

Galea, M., Shen, Q. & Levine, J. 2004. Evolutionary approaches to fuzzy modelling for classification. *The Knowledge Engineering Review* **19**(1), 27–59.

Garg, A. & Lam, J. S. L. 2015. Improving environmental sustainability by formulation of generalized power consumption models using an ensemble based multi-gene genetic programming approach. *Journal of Cleaner Production* **102**, 246–263.

Gomes, H. M., Barddal, J. P., Enembreck, F. & Bifet, A. 2017. A survey on ensemble learning for data stream classification. *ACM Computing Surveys* **50**(2), 23.

Gu, S. & Jin, Y. 2014. Generating diverse and accurate classifier ensembles using multi-objective optimization. In *Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, 9–15. IEEE.

Hansen, L. K. & Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(10), 993–1001.

Haque, M. N., Noman, M. N., Berretta, R. & Moscato, P. 2016. Optimising weights for heterogeneous ensemble of classifiers with differential evolution. In *Congress on Evolutionary Computation*, 233–240. IEEE.

Hashem, S. 1997. Optimal linear combinations of neural networks. *Neural Networks* **10**(4), 599–614.

Hauschild, M. & Pelikan, M. 2011. An introduction and survey of estimation of distribution algorithms. *Swarm and Evolutionary Computation* **1**(3), 111–128.

Hernández, L. C., Hernández, A. M., Cardoso, G. M. C. & Jiménez, Y. M. 2015. Genetic algorithms with diversity measures to build classifier systems. *Investigación Operacional* **36**(3), 206–225.

Ishibuchi, H. & Yamamoto, T. 2003. Evolutionary multiobjective optimization for generating an ensemble of fuzzy rule-based classifiers. In *Conference on Genetic and Evolutionary Computation*, 197–197. ACM.

Jackowski, K. 2014. Fixed-size ensemble classifier system evolutionarily adapted to a recurring context with an unlimited pool of classifiers. *Pattern Analysis and Applications* **17**(4), 709–724.

Jackowski, K. 2015. Adaptive splitting and selection algorithm for regression. *New Generation Computing* **33**(4), 425–448.

Jackowski, K., Krawczyk, B. & Woźniak, M. 2014. Improved adaptive splitting and selection: the hybrid training method of a classifier based on a feature space partitioning. *International Journal of Neural Systems* **24** (3), 1430007.

Joardar, S., Chatterjee, A., Bandyopadhyay, S. & Maulik, U. 2017. Multi-size patch based collaborative representation for Palm Dorsa Vein Pattern recognition by enhanced ensemble learning with modified interactive artificial bee colony algorithm. *Engineering Applications of Artificial Intelligence* **60**, 151–163.

Kaiping, L., Binglian, C., Yan, D. & Ying, H. 2013. A genetic neural network ensemble prediction model based on locally linear embedding for typhoon intensity. In *Conference on Industrial Electronics and Applications*, 137–142. IEEE.

Karakatič, S., Heričko, M. & Podgorelec, V. 2015. Weighting and sampling data for individual classifiers and bagging with genetic algorithms. In *International Joint Conference on Computational Intelligence*, 180–187. IEEE.

Kapp, M. N., Sabourin, R. & Maupin, P. 2010. Adaptive incremental learning with an ensemble of support vector machines. In *International Conference on Pattern Recognition*, 4048–4051. IEEE.

Kapp, M. N., Sabourin, R. & Maupin, P. 2011. A dynamic optimization approach for adaptive incremental learning. *International Journal of Intelligent Systems* **26**(11), 1101–1124.

Khamis, A., Xu, Y., Dong, Z. Y. & Zhang, R. 2016. Faster detection of microgrid islanding events using an adaptive ensemble classifier. *IEEE Transactions on Smart Grid* **9**(3), 1889–1899.

Kim, K.-J. & Cho, S.-B. 2005. DNA gene expression classification with ensemble classifiers optimized by speciated genetic algorithm. *In Pattern Recognition and Machine Intelligence*, **3776**, 649–653.

Kim, K.-J. & Cho, S.-B. 2008a. An evolutionary algorithm approach to optimal ensemble classifiers for DNA microarray data analysis. *IEEE Transactions on Evolutionary Computation* **12**(3), 377–388.

Kim, K.-J. & Cho, S.-B. 2008b. Evolutionary ensemble of diverse artificial neural networks using speciation. *Neurocomputing* **71**(7), 1604–1618.

Kim, K.-J. & Cho, S.-B. 2015. Meta-classifiers for high-dimensional, small sample classification for gene expression analysis. *Pattern Analysis and Applications* **18**(3), 553–569.

Kim, Y., Street, W. N. & Menczer, F. 2002. Meta-evolutionary ensembles. In *International Joint Conference on Neural Networks*, 2791–2796. IEEE.

Kiranyaz, S., Ince, T., Zabihi, M. & Ince, D. 2014. Automated patient-specific classification of long-term electroencephalography. *Journal of Biomedical Informatics* **49**, 16–31.

Kittler, J., Hatef, M., Duin, R. P. & Matas, J. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(3), 226–239.

Ko, A. H.-R., Sabourin, R. & Britto Jr., A. d. S. 2006. Evolving ensemble of classifiers in random subspace. In *Conference on Genetic and Evolutionary Computation*, 1473–1480. ACM.

Kordík, P., Černý, J. & Frýda, T. 2018. Discovering predictive ensembles for transfer learning and meta-learning. *Machine Learning* **107**, 177–207.

Kotsiantis, S. B. 2014. Bagging and boosting variants for handling classifications problems: a survey. *The Knowledge Engineering Review* **29**(1), 78–100.

Krawczyk, B., Galar, M., Jeleń, Ł. & Herrera, F. 2016. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing* **38**, 714–726.

Krawczyk, B. & Schaefer, G. 2014. Breast thermogram analysis using classifier ensembles and image symmetry features. *IEEE Systems Journal* **8**(3), 921–928.

Krawczyk, B., Schaefer, G. & Woźniak, M. 2013. A cost-sensitive ensemble classifier for breast cancer classification. In *International Symposium on Applied Computational Intelligence and Informatics*, 427–430. IEEE.

Krawczyk, B., Schaefer, G. & Woźniak, M. 2015. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. *Artificial Intelligence in Medicine* **65**(3), 219–227.

Krawczyk, B. & Woźniak, M. 2014. Evolutionary cost-sensitive ensemble for malware detection. In *SOCO/CISIS/ICEUTE*, 433–442. Springer.

Krawczyk, B., Woźniak, M. & Schaefer, G. 2014. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing* **14**, 554–562.

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J. & Woźniak, M. 2017. Ensemble learning for data stream analysis: a survey. *Information Fusion* **37**, 132–156.

Krithikaa, M. & Mallipeddi, R. 2016. Differential evolution with an ensemble of low-quality surrogates for expensive optimization problems. In *Congress on Evolutionary Computation*, 78–85. IEEE.

Krogh, A. & Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, **7**, 231–238.

Kumar, G. & Kumar, K. 2013. Design of an evolutionary approach for intrusion detection. *The Scientific World Journal* **2013**, 1–14.

Kumar, M., Husian, M., Upreti, N. & Gupta, D. 2010. Genetic algorithm: review and application. *International Journal of Information Technology* **2** (2), 451–454.

Kuncheva, L. I. & Whitaker, C. J. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* **51**(2), 181–207.

Lacy, S. E., Lones, M. A. & Smith, S. L. 2015a. A comparison of evolved linear and non-linear ensemble vote aggregators. In *Congress on Evolutionary Computation*, 758–763. IEEE.

Lacy, S. E., Lones, M. A. & Smith, S. L. 2015b. Forming classifier ensembles with multimodal evolutionary algorithms. In *Congress on Evolutionary Computation*, 723–729. IEEE.

Lévesque, J.-C., Durand, A., Gagné, C. & Sabourin, R. 2012. Multi-objective evolutionary optimization for generating ensembles of classifiers in the ROC space. In *Conference on Genetic and Evolutionary Computation*, 879–886. ACM.

Liew, W. S., Loo, C. K. & Obo, T. 2017. Optimizing FELM ensembles using GA-BIC. In *Joint World Congress of International Fuzzy Systems Association and International Conference on Soft Computing and Intelligent Systems*, 1–6. IEEE.

Lima, T. P. F. & Ludermir, T. B. 2015. Differential evolution and meta-learning for dynamic ensemble of neural network classifiers. In *International Joint Conference on Neural Networks*, 1–5. IEEE.

Liu, K., Tong, M., Xie, S. & Zeng, Z. 2014a. Fusing decision trees based on genetic programming for classification of microarray datasets. In *International Conference on Intelligent Computing*, 126–134. Springer.

Liu, K.-H., Huang, D.-S. & Zhang, J. 2007. Microarray data prediction by evolutionary classifier ensemble system. In *Congress on Evolutionary Computation*, 634–637. IEEE.

Liu, K.-H., Li, B., Zhang, J. & Du, J.-X. 2009. Ensemble component selection for improving ICA based microarray data prediction models. *Pattern Recognition* **42**(7), 1274–1283.

Liu, K.-H., Tong, M., Xie, S.-T. & Yee Ng, V. T. 2015. Genetic programming based ensemble system for microarray data classification. *Computational and Mathematical Methods in Medicine* **2015**, 1–11.

Liu, N., Cao, J., Lin, Z., Pek, P. P., Koh, Z. X. & Ong, M. E. H. 2014b. Evolutionary voting-based extreme learning machines. *Mathematical Problems in Engineering* **2014**, 1–7.

Liu, Y., Chen, W., Hu, J., Zheng, X. & Shi, Y. 2017. Ensemble of surrogates with an evolutionary multi-agent system. In *International Conference on Computer Supported Cooperative Work in Design*, 521–525. IEEE.

Lones, M. A., Smith, S. L., Alty, J. E., Lacy, S. E., Possin, K. L., Jamieson, D. R. S. & Tyrrell, A. M. 2014. Evolving classifiers to recognize the movement characteristics of Parkinson's disease patients. *IEEE Transactions on Evolutionary Computation* **18** (4), 559–576.

Ma, N., Fujita, H., Zhai, Y. & Wang, S. 2015. Ensembles of fuzzy cognitive map classifiers based on quantum computation. *Acta Polytechnica Hungarica* **12**(4), 7–26.

Mabu, S., Obayashi, M. & Kuremoto, T. 2014. Ensemble learning of rule-based evolutionary algorithm using multi layer perceptron for stock trading models. In *Joint International Conference on Soft Computing and Intelligent Systems and International Symposium on Advanced Intelligent Systems*, 624–629. IEEE.

Mabu, S., Obayashi, M. & Kuremoto, T. 2015. Ensemble learning of rule-based evolutionary algorithm using multi-layer perceptron for supporting decisions in stock trading problems. *Applied Soft Computing* **36**, 357–367.

Mauša, G. & Grbac, T. G. 2017. Co-evolutionary multi-population genetic programming for classification in software defect prediction: an empirical case study. *Applied Soft Computing* **55**, 331–351.

Mehdiyev, N., Krumeich, J., Werth, D. & Loos, P. 2015. Sensor event mining with hybrid ensemble learning and evolutionary feature subset selection model. In *International Conference on Big Data*, 2159–2168. IEEE.

Mendes-Moreira, J. A., Soares, C., Jorge, A. M. & de Sousa, J. F. 2012. Ensemble approaches for regression: a survey. *ACM Computing Surveys* **45**(1), 10:1–10:40.

Milliken, M., Bi, Y., Galway, L. & Hawe, G. 2016. Multi-objective optimization of base classifiers in stackingC by NSGA-II for intrusion detection. *In Symposium Series on Computational Intelligence*, 1–8. IEEE.

Neoh, S. C., Zhang, L., Mistry, K., Hossain, M. A., Lim, C. P., Aslam, N. & Kinghorn, P. 2015. Intelligent facial emotion recognition using a layered encoding cascade optimization model. *Applied Soft Computing* **34**, 72–93.

Obo, T., Kubota, N. & Loo, C. K. 2016. Evolutionary ensemble learning of fuzzy randomized neural network for posture recognition. In *World Automation Congress*, 1–6. IEEE.

Oehmcke, S., Heinermann, J. & Kramer, O. 2015. Analysis of diversity methods for evolutionary multi-objective ensemble classifiers. In *European Conference on the Applications of Evolutionary Computation*, 567–578. Springer.

Ojha, V. K., Abraham, A. & Snášel, V. 2017. Ensemble of heterogeneous flexible neural trees using multiobjective genetic programming. *Applied Soft Computing* **52**, 909–924.

Ojha, V. K., Jackowski, K., Abraham, A. & Snášel, V. 2014. Feature selection and ensemble of regression models for predicting the protein macromolecule dissolution profile. In *World Congress on Nature and Biologically Inspired Computing*, 121–126. IEEE.

Ojha, V. K., Jackowski, K., Abraham, A. & Snášel, V. 2015. Dimensionality reduction, and function approximation of poly (Lactic-co-glycolic acid) micro-and nanoparticle dissolution rate. *International Journal of Nanomedicine* **10**, 1119.

Olson, R. S., Bartley, N., Urbanowicz, R. J. & Moore, J. H. 2016. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Conference on Genetic and Evolutionary Computation*, 485–492. ACM.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martnez-Trinidad, J. & Kittler, J. 2010. A review of instance selection methods. *Artificial Intelligence Review* **34**(2), 133–143.

Onan, A., Korukoğlu, S. & Bulut, H. 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications* **62**, 1–16.

Opitz, D. & Maclin, R. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* **11**, 169–198.

Opitz, D. W. 1999. Feature selection for ensembles. In *National Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence Conference*, 384. American Association for Artificial Intelligence.

Oza, N. C. & Tumer, K. 2008. Classifier ensembles: select real-world applications. *Information Fusion* **9**(1), 4–20.

Pagano, C., Granger, E., Sabourin, R. & Gorodnichy, D. O. 2012. Detector ensembles for face recognition in video surveillance. In *International Joint Conference on Neural Networks*, 1–8. IEEE.

Parhizkar, E. & Abadi, M. 2015a. BeeOWA: a novel approach based on ABC algorithm and induced OWA operators for constructing one-class classifier ensembles. *Neurocomputing* **166**, 367–381.

Parhizkar, E. & Abadi, M. 2015b. OC-WAD: a one-class classifier ensemble approach for anomaly detection in web traffic. In *Iranian Conference on Electrical Engineering*, 631–636. IEEE.

Park, C. & Cho, S.-B. 2003a. Evolutionary ensemble classifier for lymphoma and colon cancer classification. In *Congress on Evolutionary Computation*, 2378–2385. IEEE.

Park, C. & Cho, S.-B. 2003b. Evolutionary computation for optimal ensemble classifier in lymphoma cancer classification. *In Foundations of Intelligent Systems*, **2871**, 521–530.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* **12**, 2825–2830.

Peimankar, A., Weddell, S. J., Jalal, T. & Lapthorn, A. C. 2016. Ensemble classifier selection using multi-objective PSO for fault diagnosis of power transformers. In *Congress on Evolutionary Computation*, 3622–3629. IEEE.

Peimankar, A., Weddell, S. J., Jalal, T. & Lapthorn, A. C. 2017. Evolutionary multi-objective fault diagnosis of power transformers. *Swarm and Evolutionary Computation* **36**, 62–75.

Pourtaheri, Z. K. & Zahiri, S. H. 2016. Ensemble classifiers with improved overfitting. In *Conference on Swarm Intelligence and Evolutionary Computation*, 93–97. IEEE.

Rahman, A. & Verma, B. 2013a. Cluster based ensemble classifier generation by joint optimization of accuracy and diversity. *International Journal of Computational Intelligence and Applications* **12**(4), 1340003.

Rahman, A. & Verma, B. 2013b. Cluster oriented ensemble classifiers using multi-objective evolutionary algorithm. In *International Joint Conference on Neural Networks*, 1–6. IEEE.

Rahman, A. & Verma, B. 2013c. Ensemble classifier generation using non-uniform layered clustering and genetic algorithm. *Knowledge-Based Systems* **43**, 30–42.

Rapakoulia, T., Theofilatos, K., Kleftogiannis, D., Likothanasis, S., Tsakalidis, A. & Mavroudi, S. 2014. EnsembleGASVR: a novel ensemble method for classifying missense single nucleotide polymorphisms. *Bioinformatics* **30**(16), 2324–2333.

Redel-Macías, M. D., Fernández-Navarro, F., Gutiérrez, P. A., Cubero-Atienza, A. J. & Hervás-Martínez, C. 2013. Ensembles of evolutionary product unit or RBF neural networks for the identification of sound for pass-by noise test in vehicles. *Neurocomputing* **109**, 56–65.

Roebber, P. J. 2015. Adaptive evolutionary programming. *Monthly Weather Review* **143**(5), 1497–1505.

Rokach, L. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* **33**(1), 1–39.

Rosales-Pérez, A., Gonzalez, J. A., Coello, C. A. C., Escalante, H. J. & Reyes-Garcia, C. A. 2014. Multi-objective model type selection. *Neurocomputing* **146**, 83–94.

Rosales-Pérez, A., García, S., Gonzalez, J. A., Coello, C. A. C. & Herrera, F. 2017. An evolutionary multi-objective model and instance selection for support vector machines with pareto-based ensembles. *IEEE Transactions on Evolutionary Computation* **21**(6), 863–877.

Sagi, O. & Rokach, L. 2018. Ensemble learning: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4), e1249.

Saha, S., Mitra, S. & Yadav, R. K. 2016. A multiobjective based automatic framework for classifying cancer-microRNA biomarkers. *Gene Reports* **4**, 91–103.

Saleh, R., Farsi, H. & Zahiri, S. H. 2016. Ensemble classification of PolSAR data using multi-objective heuristic combination rule. In *Conference on Swarm Intelligence and Evolutionary Computation*, 88–92. IEEE.

Santu, S. K. K., Rahman, M. M., Islam, M. M. & Murase, K. 2014. Towards better generalization in Pittsburgh learning classifier systems. In *Congress on Evolutionary Computation*, 1666–1673. IEEE.

Schaefer, G. 2013. Evolutionary optimisation of classifiers and classifier ensembles for cost-sensitive pattern recognition. In *International Symposium on Applied Computational Intelligence and Informatics*, 343–346. IEEE.

Schapire, R. E. 1999. A brief introduction to boosting. In *International Joint Conference on Artificial Intelligence*, 1401–1406. European Association for Artificial Intelligence.

Schuman, C. D., Birdwell, J. D. & Dean, M. E. 2014. Spatiotemporal classification using neuroscience-inspired dynamic architectures. *Procedia Computer Science* **41**, 89–97.

Shunmugapriya, P. & Kanmani, S. 2013. Optimization of stacking ensemble configurations through artificial bee colony algorithm. *Swarm and Evolutionary Computation* **12**, 24–32.

Sikdar, U. K., Ekbal, A. & Saha, S. 2012. Differential evolution based feature selection and classifier ensemble for named entity recognition. In *International Conference on Computational Linguistics*, 2475–2490. International Committee on Computational Linguistics.

Sikdar, U. K., Ekbal, A. & Saha, S. 2013. Differential evolution based mention detection for anaphora resolution. In *India Conference*, 1–6. IEEE.

Sikdar, U. K., Ekbal, A. & Saha, S. 2014a. Differential evolution based multiobjective optimization for biomedical entity extraction. In *International Conference on Advances in Computing, Communications and Informatics*, 1039–1044. IEEE.

Sikdar, U. K., Ekbal, A. & Saha, S. 2014b. Entity extraction in biochemical text using multiobjective optimization. *Computación y Sistemas* **18**(3), 591–602.

Sikdar, U. K., Ekbal, A. & Saha, S. 2015. MODE: multiobjective differential evolution for feature selection and classifier ensemble. *Soft Computing* **19**(12), 3529–3549.

Sikdar, U. K., Ekbal, A. & Saha, S. 2016. A generalized framework for anaphora resolution in Indian languages. *Knowledge-Based Systems* **109**, 147–159.

Singh, I., Sanwal, K. & Praveen, S. 2016. Breast cancer detection using two-fold genetic evolution of neural network ensembles. In *International Conference on Data Science and Engineering*, 1–6. IEEE.

Stefano, C. D., Fontanella, F., Folino, G. & Freca, A. 2011. A Bayesian approach for combining ensembles of GP classifiers. In *International Workshop on Multiple Classifier Systems*, 26–35. Springer.

Tabassum, N. & Ahmed, T. 2016. A theoretical study on classifier ensemble methods and its applications. In *International Conference on Computing for Sustainable Global Development*, 374–378. IEEE.

Tan, C. J., Lim, C. P. & Cheah, Y.-N. 2014. A multi-objective evolutionary algorithm-based ensemble optimizer for feature selection and classification with neural network models. *Neurocomputing* **125**, 217–228.

Tang, H. L., Goh, J., Peto, T., Ling, B. W.-K., Al turk, L. I., Hu, Y., Wang, S. & Saleh, G. M. 2013. The reading of components of diabetic retinopathy: an evolutionary approach for filtering normal digital fundus imaging in screening and population based studies. *PloS One* **8**(7), e66730.

Tian, J. & Feng, N. 2014. Adaptive generalized ensemble construction with feature selection and its application in recommendation. *International Journal of Computational Intelligence Systems* **7**(sup2), 35–43.

Trawiński, K., Cordón, O., Quirin, A. & Sánchez, L. 2013. Multiobjective genetic classifier selection for random oracles fuzzy rule-based classifier ensembles: how beneficial is the additional diversity? *Knowledge-Based Systems* **54**, 3–21.

Trawiński, K., Cordón, O. & Quirin, A. 2014. Embedding evolutionary multiobjective optimization into fuzzy linguistic combination method for fuzzy rule-based classifier ensembles. In *International Conference on Fuzzy Systems*, 1968–1975. IEEE.

Trivedi, S. K. & Dey, S. 2014. A study of ensemble based evolutionary classifiers for detecting unsolicited emails. In *Conference on Research in Adaptive and Convergent Systems*, 46–51. ACM.

Tsakonas, A. 2014. An analysis of accuracy-diversity trade-off for hybrid combined system with multiobjective predictor selection. *Applied Intelligence* **40**(4), 710–723.

Tsakonas, A. & Gabrys, B. 2013. A fuzzy evolutionary framework for combining ensembles. *Applied Soft Computing* **13**(4), 1800–1812.

Vaiciukynas, E., Verikas, A., Gelzinis, A., Bacauskiene, M., Kons, Z., Satt, A. & Hoory, R. 2014. Fusion of voice signal information for detection of mild laryngeal pathology. *Applied Soft Computing* **18**, 91–103.

Veeramachaneni, K., Derby, O., Sherry, D. & O'Reilly, U.-M. 2013. Learning regression ensembles with genetic programming at scale. In *Conference on Genetic and Evolutionary Computation*, 1117–1124. ACM.

Vega-Pons, S. & Ruiz-Shulcloper, J. 2011. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence* **25**(3), 337–372.

Vluymans, S., Triguero, I., Cornelis, C. & Saeys, Y. 2016. EPRENNID: an evolutionary prototype reduction based ensemble for nearest neighbor classification of imbalanced data. *Neurocomputing* **216**, 596–610.

Vukobratović, B. & Struharik, R. 2017. Hardware acceleration of nonincremental algorithms for the induction of decision trees. In *Telecommunication Forum*, 1–8. IEEE.

Wang, D. & Alhamdoosh, M. 2013. Evolutionary extreme learning machine ensembles with size control. *Neurocomputing* **102**, 98–110.

Wen, Y.-W. & Ting, C.-K. 2016. Learning ensemble of decision trees through multifactorial genetic programming. In *Congress on Evolutionary Computation*, 5293–5300. IEEE.

Winkler, S., Schaller, S., Dorfer, V., Affenzeller, M., Petz, G. & Karpowicz, M. 2015. Data-based prediction of sentiments using heterogeneous model ensembles. *Soft Computing* **19**(12), 3401–3412.

Wistuba, M., Schilling, N. & Schmidt-Thieme, L. 2017. Automatic frankensteining: creating complex ensembles autonomously. In *International Conference on Data Mining*, 741–749. SIAM.

Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* **5**(2), 241–259.

Wolpert, D. H. & Macready, W. G. 1997. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1** (1), 67–82.

Woon, W. L. & Kramer, O. 2016. Enhanced SVR ensembles for wind power prediction. In *International Joint Conference on Neural Networks*, 2743–2748. IEEE.

Wozniak, M. 2009. Evolutionary approach to produce classifier ensemble based on weighted voting. In *World Congress on Nature and Biologically Inspired Computing*, 648–653. IEEE.

Xavier-Júnior, J. A. C., Freitas, A. A., Feitosa-Neto, A. & Ludermir, T. B. 2018. A novel evolutionary algorithm for automated machine learning focusing on classifier ensembles. In *Brazilian Conference on Intelligent Systems*, São Paulo, Brazil. IEEE.

Xu, H., Caramanis, C. & Mannor, S. 2012. Sparse algorithms are not stable: a no-free-lunch theorem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(1), 187–193.

Yager, R. R. 1988. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on systems, Man, and Cybernetics* **18**(1), 183–190.

Yao, X. & Islam, M. M. 2008. Evolving artificial neural network ensembles. *IEEE Computational Intelligence Magazine* **3**(1), 31–42.

Zagorecki, A. 2014. Feature selection for Naive Bayesian network ensemble using evolutionary algorithms. In *Federated Conference on Computer Science and Information Systems*, 381–385. IEEE.

Zhang, L., Mistry, K., Neoh, S. C. & Lim, C. P. 2016a. Intelligent facial emotion recognition using moth-firefly optimization. *Knowledge-Based Systems* **111**, 248–267.

Zhang, W., Qu, Z., Zhang, K., Mao, W., Ma, Y., Fan, X. 2017a. A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. *Energy Conversion and Management* **136**, 439–451.

Zhang, Y., Liu, B. & Yang, F. 2016b. Differential evolution based selective ensemble of extreme learning machine. In *Trustcom/BigDataSE/ISPA*, 1327–1333. IEEE.

Zhang, Y., Liu, B., Cai, J. & Zhang, S. 2017b. Ensemble weighted extreme learning machine for imbalanced data classification based on differential evolution. *Neural Computing and Applications* **28**(1), 259–267.

Zhang, Y., Zhang, H., Cai, J. & Yang, B. 2014. A weighted voting classifier based on differential evolution. *Abstract and Applied Analysis* **2014**(1), 1–6.