The Journal of Systems & Software 195 (2023) 111542

Contents lists available at ScienceDirect

The Journal of Systems & Software

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

Ensemble Effort Estimation: An updated and extended systematic literature review[☆]

José Thiago H. de A. Cabral, Adriano L.I. Oliveira [*], Fabio Q.B. da Silva

Centro de Informática (CIn), Universidade Federal de Pernambuco, Recife-PE, Brazil

ARTICLE INFO

Article history:
Received 9 November 2021
Received in revised form 7 October 2022
Accepted 10 October 2022
Available online 13 October 2022

Keywords:
Systematic literature review, Ensemble Effort Estimation
Software Engineering
Software effort estimation
Machine learning

ABSTRACT

Ensemble Effort Estimation (EEE) techniques combine several individual software estimation methods in order to address the weaknesses of individual methods for prediction tasks. A systematic review published in 2016 analyzed empirical studies on EEE techniques published between 2010 and (January) 2016. The research on EEE has continuously evolved over the past five years (2016–2020), generating new findings that should be aggregated to the existing body of evidence on the subject.

The goal of this paper is to update the systematic review from 2016 with new findings from studies published between 2016 (full year) and 2020 (inclusive).

To conduct our review update, we followed existing guidelines for updating systematic reviews in software engineering and other fields. We started with an appraisal of the background and methods of the 2016 review, which resulted in the updated review protocol used to conduct our study. We retrieved 3,682 papers using automatic searching techniques, from which we selected 30 papers for data extraction and analysis.

Our findings reinforce the results of the previous review in that machine learning is still the technique most common to construct EEE and that the ensemble techniques have outperformed the individual models. We added new evidence showing that there is no clear superiority of an EEE model over the others. Also, we found that ensemble dynamic selection is still little used in Software Effort Estimation (SEE).

This review adds new evidence about the use of EEE techniques in software development which reinforces previous findings and also shows research opportunities in constructing more effective EEE. Besides, ensemble dynamic selection appears as a promising area of research which still is underexplored.

1. Introduction

Software Development Effort Estimation (SDEE) is well known to be one of the main challenges in software development (Trendowicz and Jeffery, 2014). Many projects have failed due to under or over effort and/or cost estimation. On the one hand, estimations that exceed the development deadline tend to lead the clients to give up on sponsoring the project. On the other hand, overly optimistic estimations can lead to an increase in resources, mostly human and financial, which at times preclude the development of the product. In that sense, the main reasons for entailing flaws in software development are mainly linked to unrealistic goals, weak resource estimations that are vital to develop the project, and the incapacity to better understand the complexity of the project (Charette, 2005).

A systematic review published in 2012, presented various models used to predict the software effort estimation (Wen et al., 2012). The authors concluded that the use of machine learning models, such as Artificial Neural Networks, had increased in the effort estimation literature. Moreover, machine learning techniques had, for the most part, shown better accuracy than those that do not make use of learning-based techniques. Even though many such techniques exist, a common ground has not been found in order to determine, in general terms, the best technique (Wen et al., 2012), because the performance of each technique varies from dataset to dataset. This leads us to believe that the development of a single technique capable of being superior within all software estimation effort contexts is not feasible.

A more recent approach that combines individual methods, known as Ensemble Effort Estimation (EEE), has been used to try

[☆] Editor: Burak Turhan.

[*] Correspondence to: Centro de Informática (CIn), Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, s/n, Cidade Universitária (Campus Recife), CEP: 50.740-560 Recife-PE, Brazil.

E-mail addresses: jthac@cin.ufpe.br (J.T.H.d.A. Cabral), alio@cin.ufpe.br (A.L.I. Oliveira), fabio@cin.ufpe.br (F.Q.B. da Silva).

https://doi.org/10.1016/j.jss.2022.111542
0164-1212/© 2022 Published by Elsevier Inc.

to overcome the limitations of single techniques. The effort predicted by an EEE technique is achieved from the (linear or nonlinear) combination of the results of individual methods. According to Dietterich (2000), the EEE techniques can be grouped into two categories: homogeneous and heterogeneous. A homogeneous ensemble is created from an individual technique that has at least two different configurations, whereas a heterogeneous ensemble is built from distinct individual techniques.

Idri et al. (2016) conducted a systematic review to analyze the use of EEE in SDEE, hereafter referred to as the 2016 Review. Idri et al. analyzed 24 empirical studies published between 2000 and (January) 2016, and their main findings can be summarized as follows:

- Consistently with Dietterich (2000), EEE techniques may be separated into two types: homogeneous and heterogeneous.
- Machine learning single models are the most frequently employed in constructing EEE techniques, which is consistent with the findings from Wen et al. (2012) for single models.
- EEE techniques usually yielded acceptable estimation accuracy, although the results came from different evaluation designs, making it difficult to "define the conditions under which the results were obtained".
- Twenty rules to combine the estimates in the EEE techniques were found and classified as linear and nonlinear combination rules.
- EEE techniques are usually more accurate than single models. However, the researchers emphasize that the sample of 25 studies was small and the experiments were conducted under different conditions, making it difficult to integrate and interpret the results.
- Homogeneous ensembles seemed to be more accurate, but also the number of studies were small to support stronger conclusions.
- Four techniques for selecting constituent methods in the ensemble were found. However, few studies "clearly explained the methodology used to select ensemble constituents".

The 2016 Review provided an important conceptual framework for understanding the landscape of the research on EEE up to 2016. However, the small number of studies and the use of different experimental designs made it difficult to integrate and interpret the results towards building stronger empirical evidence. The researchers suggested that more studies would be needed, hopefully using comparable experimental designs, in order to support more conclusive results on the topic.

The research questions asked in the 2016 Review are still relevant nowadays, both for academia and industry. Moreover, the research on EEE has continuously evolved over the past six years (2016–2020), generating new findings that should be compared and integrated with those from the 2016 Review. Therefore, the goal of this paper is to update the 2016 Review with new findings from studies published between 2016 (full year) and 2020 (inclusive).

To conduct the review update, we followed the guidelines for updating systematic reviews first proposed by Garner et al. (2016) and also applied by Mendes et al. (2020) in the software engineering context. Consistently with Garner et al. (2016), we performed the following steps, summarized here and explained in more details in Section 3:

- We used the decision framework of Garner et al. (2016) to guide and support our decision for updating the 2016 Review.
- We refreshed background, research questions, inclusion criteria, and methods, so as to create an updated review protocol.
- We updated the search process according to the changes in background and research questions and also to account for changes in the used search engines.
- We finally conducted the update following the updated protocol and search process.

The rest of this paper is organized as follows. In Section 2, we present the research method used in our work, including the comparison between the 2016 Review and the review update. In Section 3, we present the results of the review update. In Section 4, we discuss our results and the potential limitations and threats to validity. Finally, in Section 5, we present conclusions and directions for future work.

2. Review method

In this section, we present the review method used to guide our review update. We followed the guidelines proposed by Kitchenham and Chalmers for conducting systematic reviews in software engineering (Kitchenham and Charters, 2006), also the guidelines proposed by Garner et al. (2016) for updating systematic reviews and the snowballing search process guidelines proposed by Wohlin (2014). This section is then structured as follows. We first show the application of the decision framework from Garner et al. (2016) to support our decision to update the 2016 Review. We then describe the refreshed review protocol that was constructed to guide the review update, which also compares the two reviews regarding objectives and research questions. The replication package composed of the review protocol and the raw data set produced by the data extraction process can be found at (https://drive.google.com/file/d/1TCUlUnPibkbP-W9jZaqY_JU7apSzxDKS/view?usp=sharing).

2.1. Deciding to update the 2016 review

The decision framework proposed by Garner et al. (2016) is composed of a series of questions to guide the decision process and a set of standard terms to be used to report the decision. In Table 1, we present the questions, our answers and their justifications that lead to the decision of updating the 2016 Review.

According to the guidelines, our assessment of the 2016 Review and of the current state of the art (in 2020) supported our decision to update the review.

We then checked if our review would fit the definition of review update used by Garner et al. (2016):

"... an update of a systematic review as a new edition of a published systematic review with changes that can include new data, new methods, or new analyses to the previous edition. … An update asks a similar question … and has similar objectives; thus it has similar inclusion criteria. These inclusion criteria can be modified in the light of developments within the topic area ... Updates will include a new search for potentially relevant studies and incorporate any eligible studies or data; and adjust the findings and conclusions as appropriate".

Our review fits the above definition as it includes new data and a refreshed review protocol that asks similar questions, has similar inclusion criteria and a new search process that incorporates new relevant studies. Also, we integrated and adjusted the findings in light of the new evidence

2.2. Review protocol refresh

After deciding to update the 2016 Review, we followed (Garner et al., 2016) guidelines to refresh the review protocol. We structured the review method according to the guidelines of Kitchenham and Charters (2006), which proposed the steps illustrated in Fig. 1.

Table 1
Questions, answers and justifications about updating the 2016 review.

Questions (extracted from Garner et al. (2016))	Answers	Justifications
Does published review still address a current question?	Yes	The study of EEE in the context of SDEE is still relevant and of practical as well as academic interest.
Has review had good access or use?	Partially	The 2016 Review has over 100 citations on Google Scholar, showing that it has been used mostly in academic settings.
Review used valid methods and was well conducted?	Yes	The review followed established guidelines and has been published in a high quality peer reviewed journal.
Are there any new relevant methods?	Yes	We performed an initial non systematic search and found that dynamic ensemble selection methods have been applied recently.
Are there any new studies, or new information?	Yes	We performed an initial non systematic search and found new studies published in high quality venues.
Will adoption of new methods change findings or credibility?	Maybe	Given the small number of studies analyzed in the 2016 Review, it is likely that new methods will add new findings to the area.
Will new studies/information/data change findings or credibility?	Maybe	Given the small number of studies and the different study designs analyzed in the 2016 Review, it is likely that new studies, information and/or data will add new findings to the area.

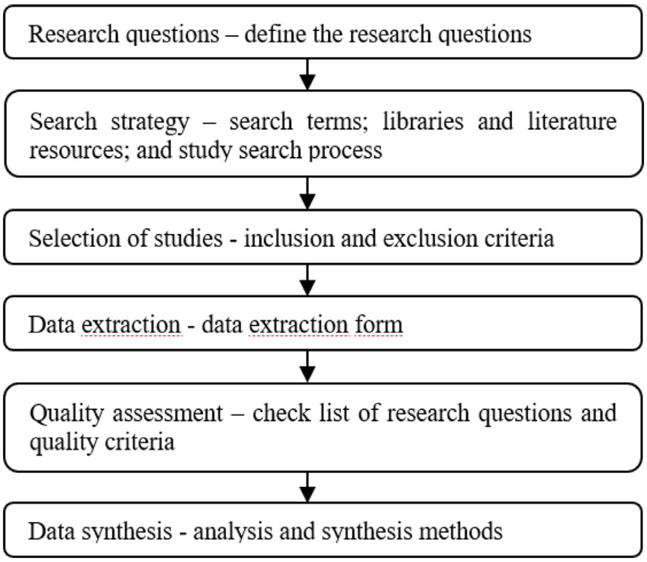

Research questions – define the research questions
Search strategy – search terms; libraries and literature resources; and study search process
Selection of studies - inclusion and exclusion criteria
Data extraction - data extraction form
Quality assessment – check list of research questions and quality criteria
Data synthesis - analysis and synthesis methods

Fig. 1. Stages of the review process (Kitchenham and Charters, 2006).

2.2.1. Research questions

We construct our research questions by refining the six questions from the 2016 Review. In Table 2, we present the questions from the 2016 Review, our questions and the explanations for the changes we made in the review update. Overall, the refreshed research questions were formulated to allow: (i) to update the results from 2016 Review; (ii) to investigate the trend overtime in the use of EEE in SDEE; (iii) and finally, to investigate the use of dynamic ensemble selection methods in SDEE, which was not addressed in the 2016 Review.

2.2.2. Search strategy

We used a mix of automatic and manual search to increase coverage. For the automatic search, we started with the search string from the 2016 Review and added terms to filter type and year of publication (Table 3). We only searched for articles published in conference proceedings and journals, and restricted the publication time for interval 2016–2020 (inclusive).

Automatic search was performed on three search engines: Scopus, ACM and IEEE. The search considered the title, the abstract, and the keywords. After removing duplicates, the retrieved articles were pre-selected based on the contents of abstract, title and keywords. Pre-selected articles were further analyzed applying the inclusion and exclusion criteria, which are presented and explained in Section 2.2.3.

After applying the inclusion and exclusion criteria on the results of the automatic search, the set of selected articles was used as the start set for the snowballing search process, for which we followed the guidelines proposed by Wohlin (2014). As in the automatic search, after a pre-selection of the articles found during snowballing search, we applied the inclusion and exclusion criteria to select the final articles in the manual search process.

2.2.3. Selection process

This step aims to select the studies that are relevant to answer the research questions. We applied the following inclusion and exclusion criteria to the articles pre-selected from the automatic and snowball searches, described in the previous section.

(a) Inclusion criteria — we included only peer reviewed articles, published in journals or in proceedings of scientific events, which met at least one of the conditions below:

- Studies that used or compared different EEE techniques or EEE versus individual techniques.
- Studies that used or compared Dynamic Ensemble Selection (DES) techniques applied to software effort estimation.
- Studies that used hybrid models, but with an EEE approach.

(b) Exclusion criteria — we excluded studies that met at least one of the criteria below:

- Studies whose goals were not related to the effort estimation or software cost;
- Hybrid studies that did not consider EEE in the proposed version;
- Studies that considered a single technique of software effort estimation.

To apply the inclusion and exclusion criteria, each pre-selected article was read in detail by two researchers, each one working independently (first and second authors). An article would be selected if it met at least one inclusion criterion and none of the exclusion criteria. Disagreements between the two researchers were solved in a consensus meeting. The list of selected studies is shown in Table 20 of Appendix B, and the list of studies that answer the research questions (RQ1–RQ6) is presented in Table 4.

Table 2
Comparing research questions.

2016 review (Idri et al., 2016)	The review update	Remarks and explanations
2016-RQ1 What are the most frequently used EEE techniques?	RQ1 What are the most frequently used EEE techniques (homogeneous/heterogeneous)? RQ2 What are the most frequently used individual techniques in the composition of EEE?	We split the 2016-RQ1 into our RQ1 and RQ2 to make it easier to structure and discuss the findings.
2016-RQ2 What is the overall estimation accuracy of EEE techniques?		Not addressed because the metric used in the original review has been found to be biased and may favor some types of methods over others (Shepperd and MacDonell, 2012).
2016-RQ3 What are the different rules used to combine estimation models?	RQ3 What are the different evaluation metrics, databases, matching rules, validation methods and statistical testing used in EEE?	Our RQ3 includes the 2016-RQ3 and adds further information.
2016-RQ4 Do EEE techniques outperform the single models used in SDEE?	RQ4 Do EEE techniques outperform the single models used in SDEE?	Same questions. We focused on the performance of the learners.
2016-RQ5 Are there some EEE techniques that clearly outperform others?	RQ5 Are there some EEE techniques that clearly outperform others, regarding type, rules and composition?	Our RQ5 includes the 2016-RQ5 and adds further comparisons.
	RQ6 Has dynamic model selection been used in the EEE literature?	Dynamic model selection is relatively new and has been used in other applications. We decided to check its application in EEE.
2016-RQ6 How to construct ensembles?		This question was not addressed.

Table 3
Search string.

Software AND (effort OR cost*) AND (estimat* OR predict* OR assess* OR forestcast*) AND (ensemble OR taxonomy OR committee OR sampling OR fusion OR boosting OR bagging OR random OR multipl* OR combin* OR cluster* OR bootstrap OR meta*)) AND DOCTYPE (ar OR cp) AND PUBYEAR > 2015 AND PUBYEAR < 2021

Table 4
Relationship of studies with research questions.

Research question	Number of studies	Selected studies
RQ1	25	S1; S3–S21; S23–S24; S26–S27; S29
RQ2	28	S1–S2; S4–S13; S30
RQ3	24	S1–S4; S6–S11; S15–S16; S18; S20–S21; S24; S27; S29
RQ4	27	S1–S8; S10–S15; S17–S28; S30
RQ5	19	S3–S11; S13; S15–S16; S18; S20–S21; S23–S24; S29–S30
RQ6	2	S5; S8

2.2.4. Data extraction

The data extraction process consisted of three steps, which extracted data from each article in an incremental way:

1. First, the descriptive and reference data were extracted directly from the title page of each article.
2. Second, we performed a primary data extraction after a full reading of each article using the fields described in Table 18 (Appendix A). This allowed us to get familiarized with the data available in the articles and to prepare for the next round of data extraction.
3. Third, we expanded and refined the attributes used to guide data extraction based on the results of the primary data extraction, creating the fields described in Table 19 (Appendix A). We then reviewed each article incrementally once more to complement that primary data extraction with the new attributes.

Data extraction was performed by the first author. The data extracted was checked by the second author and disagreements were resolved in a consensus meeting. The third author checked the data extraction on a random sample covering 50% of the studies and no disagreements were found at this point.

2.2.5. Quality assessment

In parallel with the data extraction, eight quality assessment questions were applied to the selected studies to improve the selection criteria. Quality assessment is paramount in order to synthesize and interpret the findings from the collected data (Kitchenham and Charters, 2006). We used almost the same set of criteria used in the 2016 Review. We dropped their QA8 (Does the study make a contribution to academia or the industry community?) because we considered it too difficult to find an operationalization that would result in a consistent and unbiased assessment. Table 5 presents the quality assessment questions used in the review update and the assessment criteria. These questions were answered as "No", "Yes", or "Partial", and scored zero (0), one (+1.0) and half (+0.5), respectively. Table 6 shows the scores of the studies by quality assessment question.

The overall quality score of each study was computed as the mean of the individual score for each QA. The quality assessment process was carried out by 2 researchers (the first and the second author of this paper) that applied the same criterion for all the studies, previously defined. The column assessment criteria in Table 5 presents these criteria in detail. We used the threshold of 0.5 as the minimum quality score for a study to remain in the selected set. No study scored less than 0.5. The details of the quality score of the studies are shown in Table 6.

2.2.6. Data synthesis

Data synthesis aims to aggregate evidence for each research question. Aggregation may increase the strength of the evidence when the individual studies agree on their findings. On the other hand, evidence pointing to different directions indicates inconclusive results and may weaken the evidence of some, or all individual findings.

Table 5
Quality assessment questions.

Id	Questions	Assessment criteria
QA1	Were the objectives presented in the study clearly defined?	The averages of the researchers' assessments, rounded.
QA2	Was the proposed solution clearly well defined and discussed?	It was considered whether the authors presented in the study: • Discussion of specific features of the proposal; • Tables, graphs and statistical tests for comparison of results.
QA3	Is there more than one solution proposed and discussed?	• More than one version of the proposal was discussed = score (1.0) • A single version was well discussed = score (0.5) • One version without discussion = 0 (score)
QA4	Was the experiment carried out applied to a sufficient number of databases?	• More than 5 datasets = score (1.0) • Between 3 and 5 datasets = score (0.5) • Less than 3 datasets = score (0)
QA5	Was the measured accuracy well interpreted?	We considered the number of metrics: • More than 2 metrics well interpreted = score (1.0); • 1 metric well interpreted; • 1 metric with weak interpretation.
QA6	Has the proposed solution been compared with other techniques?	We considered the number of techniques and if it was used in statistical tests. • More than 1 technique and statistical tests = score (1.0); • 1 technique with statistical tests = score (0.5); • 1 technique without statistical tests = score (0)
QA7	Was the estimation context well defined?	If the objectives, proposal and results were well contextualized in the Introduction section
QA8	Were the findings clearly presented and supported by the results obtained?	We considered: • Whether the results were presented in tables, graphs and with statistical tests; • Whether there is a section in the study that discussed the findings; • If the results were consistent with the discussion presented.

Table 6
Evaluation of the quality of studies.

Study id	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	Mean
S1	1	1	0	0	1	0	1	0.5	0.6
S2	1	1	0	0.5	1	0.5	1	0.5	0.7
S3	1	1	0.5	0	1	0.5	1	0.5	0.7
S4	1	1	1	0.5	1	0.5	1	0.5	0.8
S5	1	1	1	1	0.5	0.5	1	0.5	0.8
S6	1	0.5	1	1	0.5	0.5	0.5	0.5	0.7
S7	1	1	1	0	1	0.5	1	0.5	0.8
S8	1	0.5	1	1	0	1	1	0.5	0.8
S9	1	1	1	0.5	1	0.5	1	0.5	0.8
S10	1	1	1	0.5	1	0.5	1	0.5	0.8
S11	1	1	1	0.5	0.5	0.5	1	0.5	0.8
S12	1	1	0	0	1	0	1	0.5	0.6
S13	1	1	0	0.5	1	0	1	0.5	0.6
S14	1	0.5	0	0	1	0	1	0.5	0.5
S15	1	1	1	0.5	1	0.5	1	0.5	0.8
S16	1	1	0.5	0.5	1	0.5	1	0.5	0.8
S17	1	1	0	0	1	0.5	1	0.5	0.6
S18	1	0.5	0.5	0.5	1	1	0.5	0.5	0.7
S19	1	1	0	0.5	1	0	1	1	0.7
S20	1	1	0.5	0	1	1	1	1	0.8
S21	1	0.5	0.5	0	0.5	0.5	1	0.5	0.6
S22	1	1	0	0.5	1	0.5	1	1	0.8
S23	1	1	1	1	0.5	1	1	1	0.9
S24	1	1	1	1	1	1	1	0.5	0.9
S25	1	1	0	0.5	0.5	0.5	1	0.5	0.6
S26	1	1	0	0	1	1	1	0.5	0.7
S27	1	1	1	0	0.5	0.5	1	1	0.8
S28	1	1	0	1	1	0.5	1	1	0.8
S29	1	1	0.5	0	1	0.5	1	0.5	0.7
S30	1	1	0.5	0.5	1	0.5	1	0.5	0.8

We used the same methods of synthesis applied in the 2016 Review:

- We aggregated evidence for RQ1, RQ2, RQ3 and RQ6. We tabulated the data for each research question and then used visualization tools to display the aggregated results.
- Vote counting was used to support answers to RQ4 and RQ5. It consisted in summing up the number of studies in which one model outperformed other models.

Table 7
Number of articles from each source.

	Search result	Non-repeated	Pre-selected	Selected
Scopus	1902	1899	146	17
ACM	622	453	14	2
IEEE	1158	1056	30	5
Snowballing	0	0	0	6
Total	3682	3408	190	30

3. Results

In this section, we present the results of the review update and compare our findings with the 2016 Review, whenever appropriate. We start with descriptive information about the new set of reviewed studies. Then, we present the synthesis of evidence for research questions RQ1–RQ6, presented in Section 2.2.1.

3.1. Descriptive information about selected studies

Fig. 2 summarizes the search and article selection processes, and Table 7 presents the number of studies found per library and also from the snowballing process.

The automatic search retrieved 3682 articles from three databases. After removing duplicates (same article retrieved from more than one database), 3408 unique articles remained. A pre-selection of these studies was carried out by reading the title, keywords and the abstract. The pre-selection list consisted of 190 studies that seemed to be relevant or that raised doubts regarding their importance. We then applied inclusion and exclusion criteria to the 190 pre-selected studies, resulting in the selection of 24 articles. We then used these 24 articles as the start set for a snowball search, in which we checked the references and citation of the 24 articles manually. Snowballing search produced 6 new

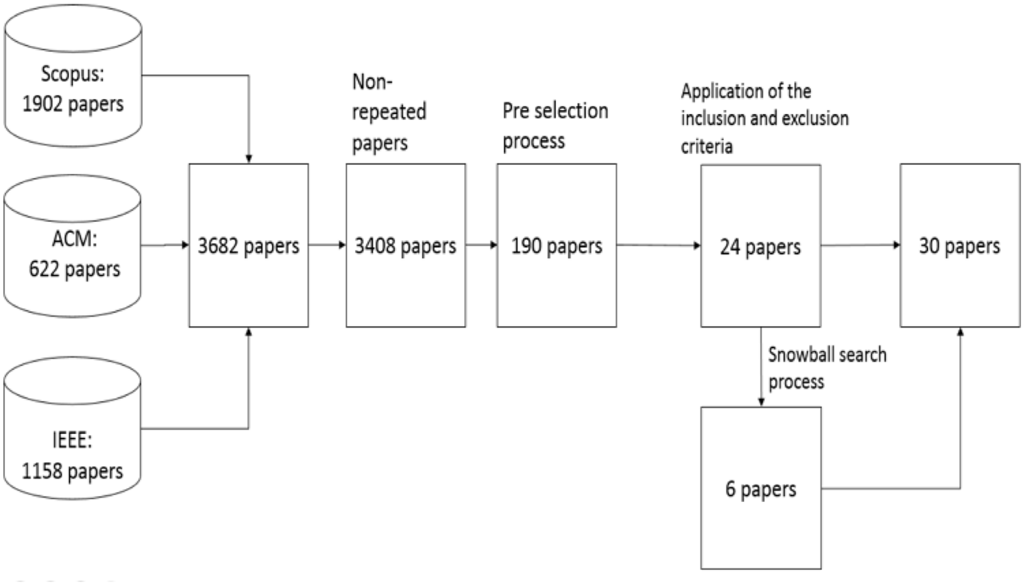


Fig. 2. Selection process of relevant studies.

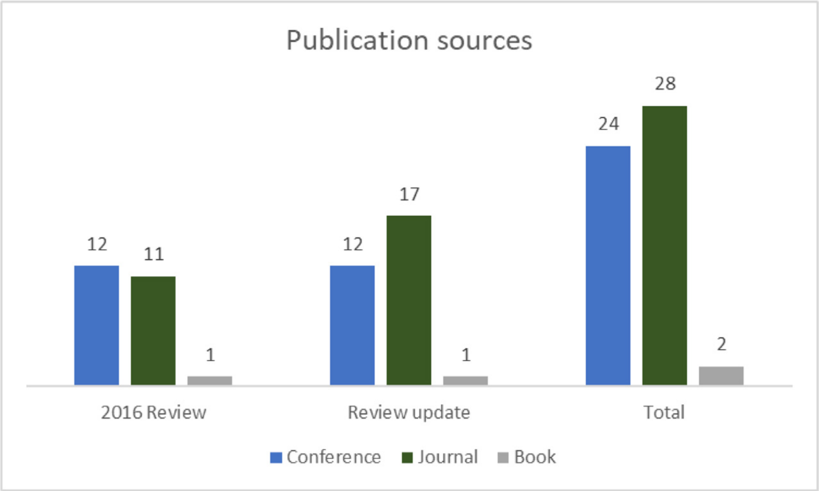


Fig. 3. Number of studies by source from 2006 to 2020.

Table 8
Publication sources (conferences and journals as in Table 5 of the 2016 Review).

	2016 review	% (N = 24)	Review update	% (N = 30)	2000–2020	% (N = 54)
Conference	12	50%	12	40%	24	44%
Journal	11	46%	17	47%	28	52%
Book chapter	1	4%	1	3%	2	4%
Total	24	100%	30	100%	54	100%

relevant articles, resulting in the final selection of 30 articles. Table 8 presents the distribution of articles between publication sources.

Table 8 also shows the percentage of articles derived from conferences, journals, and book chapters. There is a balance between conference and journal publications. Nevertheless, we can see that the participation of conference papers dropped 10% between the 2016 Review and the review updated. Although we do not have empirical evidence, we may hypothesize that the increase in the number of journal articles reflects an increase in maturity of the studies. Fig. 3 illustrates this increase.

At first, the studies were gathered taking into consideration the year. By analyzing the results, it is easily observed (Fig. 4) that there is a behavior pattern in terms of the number of studies that discuss EEE throughout the years. The quantity of studies tends to increase in some years; in Fig. 4, we can see it in years 2009, 2013 and 2018. These increases in studies in the area may be due to the emergence of some machine learning techniques, and then there is a decrease in the following years until the number of studies increase once again. It is easy to identify the increase in certain years. However, the density of studies from 2016 up until the end of 2020 is higher when compared to the 2016 Review. A greater number of studies were published in the past five years than from 2006 to 2015 (30 vs. 25 studies). This leads us to believe that EEE will continue to be discussed extensively in the literature, and that new studies might emerge in the next few years. We also believe that deep learning and dynamic selection of multiple classifiers will be the most researched approaches subsequently,

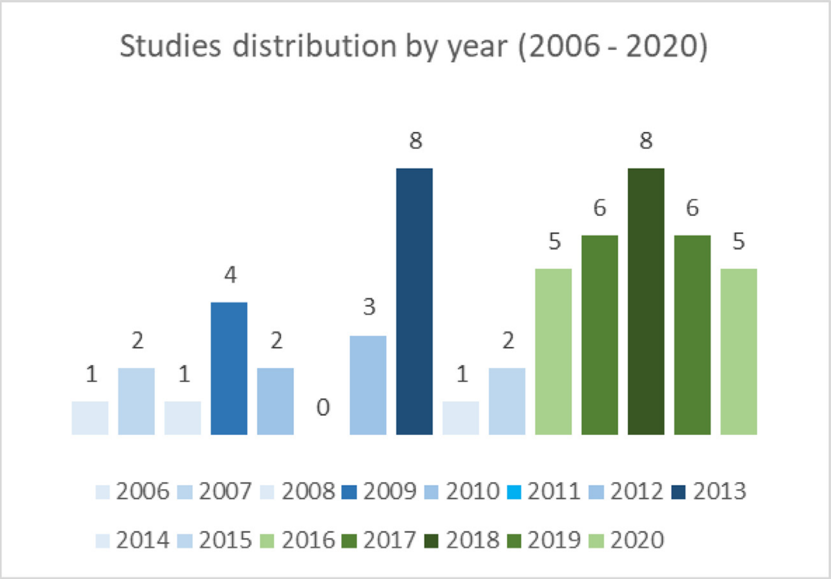


Fig. 4. Temporal distribution of selected studies (2016 Review and Updated Review).

Table 9
Most frequent researchers involved in EEE studies from 2006 to 2020.

Author	Number of papers
Ali Idri	9
Alain Abran; Mohamed Hosni	7
Ali Bou Nassif	6
Le Thi My Hanh; Leandro Minku	5
Xin Yao	4
O. Malgonde; Mahmoud O. Elish; Mohammad Azzeh; Namir Abdelwahed	3
Ayse Bener; Abdelali Zakrani; Adriano L. I. Oliveira; Cuauhtémoc López Martín; Danny Ho; Ekrem Kocaguneli; Emília Mendes; Gultekin, M.; Jarley P. Nobrega; L. F. Capretz; Lefteris Angelis; Liang Song; M. Carr; Nikolaos Mittas; Palaniswamy, S.K.; Przemyslaw Pospieszny; Santanu Kumar Rath; Shashank Mouli Satapathy; Tim Menzies; V. Ravi; Y. Kultur	2

since these approaches have been given little attention until this moment.

Nevertheless, it is fundamental that new researchers are interested in EEE so that the quantity of studies continues to increase, because many studies that were selected were carried out by the same authors. Table 9 shows the amount of studies per author, considering the result presented in Idri et al. (2016) too. In this work, we do not take into account if the author was the main author. Table 9 shows only authors with at least 2 articles. A total of 114 authors with a least one paper were identified in the studies from 2006 to 2020.

3.2. RQ1 - What are the most frequently used EEE techniques?

The main motivation for RQ1 is to identify the relative frequency of use of homogeneous and heterogeneous EEE. Homogeneous ensemble models from a single individual technique, but with different configurations of the selected technique. Heterogeneous ensembles combine distinct techniques. The rules used for the combination process do not influence the definition of the types of ensembles.

In the 2016 Review, 58% (14/24) of the articles discussed only homogeneous ensembles, nearly 25% (9/24) of the articles investigated only heterogeneous ensembles, and just over 1% (3/24) of the articles investigated both types. In the review update, the relative frequency of use of the two types of ensembles is more uniform, showing an increase in the use of heterogeneous ensembles. Just under 50% (14/30) articles investigated only homogeneous ensembles, 40% (12/30) of the articles investigated only heterogeneous ensembles, and just over 1% (4/30) of the articles investigated heterogeneous and homogeneous ensembles. Fig. 5 illustrates the relative frequency of use of both types of ensembles, where we added the number of articles investigating both types of ensembles to both columns.

The increase in studies with heterogeneous ensembles reached 78% from 2016 to 2020, whereas the quantity of studies with homogeneous ensembles have almost not increased around the same period of time. However, the homogeneous ensembles are still superior in terms of the overall quantity of studies.

We also analyzed the methods for generating models. Fig. 6 shows that Bagging, Boosting and Random Forest are found to be the most widely used methods. It presents the results considering the total number of studies (2006–2020). Also, the use of Boosting and Random Forest has greatly increased in the past few years. Out of the 9 studies collected, 7 used Boosting in the past 5 year and 7 used Random Forest in the past 3 years, thus Boosting and Random Forest increased 250%. The Bagging method is still widely used, even though it has not increased when compared to its main competitors. Bagging increased 40%, 7 of the 12 studies collected, in relation to the 2016 Review.

In general, heterogeneous ensembles are formed by the combination of distinct methods, and such combinations are discussed in another research question. However, Stacking is a technique for creating heterogeneous ensembles, which were discussed in 8 studies out of 25.

It is important to note that a given article can present many different ensembles. Because of that and the results achieved, we can state that the use of heterogeneous ensembles has recently increased in the literature in EEE, and the creation of

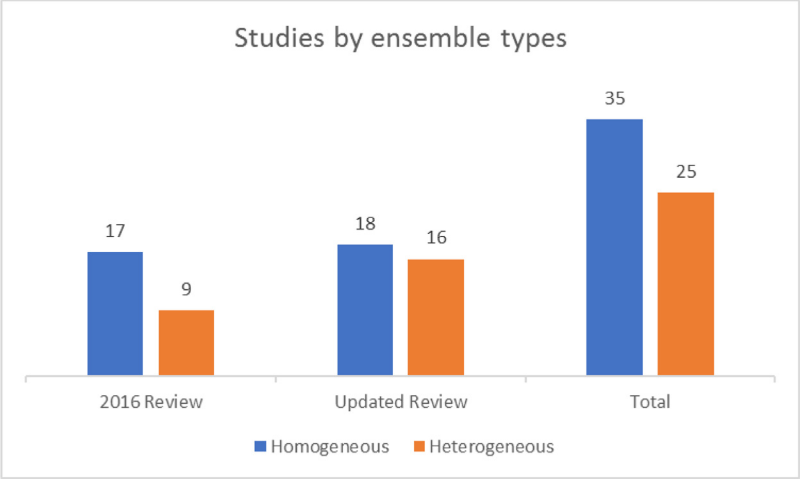


Fig. 5. Number of studies by ensemble types.

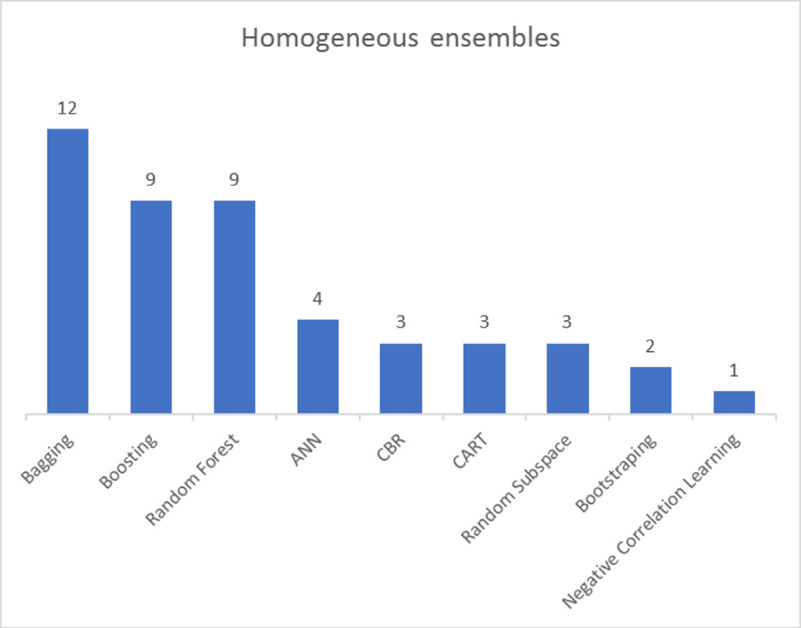


Fig. 6. Number of homogeneous ensembles by generation methods from 2006 to 2020.

homogeneous models has remained consistent mainly due to the contribution of works that adopted the Random Forest method.

3.3. RQ2 - What are the most frequently used individual techniques in the composition of EEE?

In this section, we discuss the frequency of use of individual techniques in order to create EEE models. For the purpose of counting the frequency of these individual techniques in homogeneous and heterogeneous ensembles, the following procedure was adopted:

- If the technique is heterogeneous, each base model is counted once;
- If the technique is homogeneous, the base model is counted once.

Fig. 7 shows that the neural networks and the regression trees are the most commonly used base methods in the homogeneous ensembles. Various types of neural networks were considered; Multilayer Perceptron (MLP) was used more frequently, and also different types of trees, both classification and regression trees. The algorithms used in the studies were grouped so as to facilitate the presentation. For example, Decision Tree, Regression Tree, or Classification and Regression Trees were considered tree models.

The Cased Based Regression (CBR) is the third most widely used technique in homogeneous ensembles after the neural networks and the decision trees. The ease in understanding the algorithm, the possibility of different hyperparameter values and the number of studies contributed to the use of CBR. These results are consistent between the 2016 Review and the Updated Review, as can be seen in Fig. 7.

As for the heterogeneous ensembles, Fig. 8 presents the number of studies that used each one of the groups of algorithms under analysis. Artificial Neural Networks (ANN) were the most widely used grouping algorithm and MLP was the most commonly used type of ANN. Moreover, CBR, Regression Trees (RT), Linear Regression (LR) and Support Vector Regression (SVR) methods were the individual techniques that showed the largest increase in terms of the frequency of use when compared to the

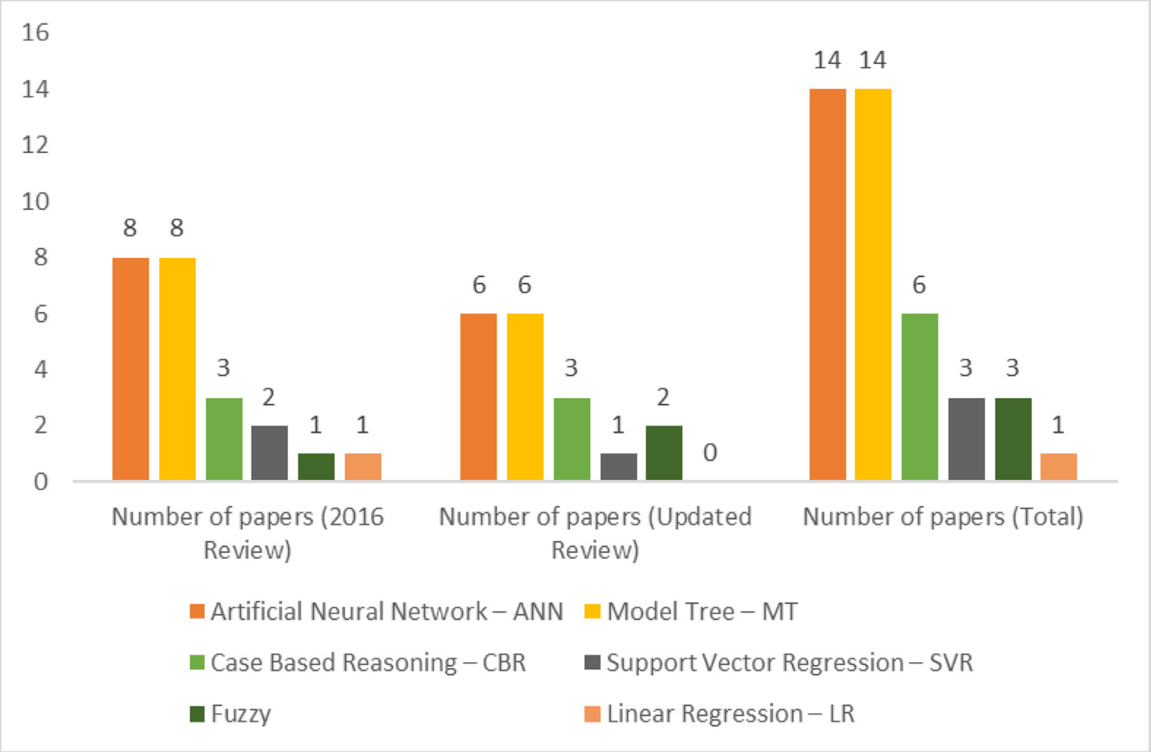


Fig. 7. Distribution of basic methods used in homogeneous ensembles by studies.

previous work (140%, 175%, 300% and 133%, respectively), while ANN increased 20%, from 10 to 12 studies. In the updated review, the traditional (parametric) techniques, which are more widely known in the industry, were also mentioned.

The following methods are emphasized because they are simple and do not require the use of ML algorithms: Cocomo, Function Points and Slim. For the heterogeneous structure, the CBR was more widely used than the regression trees. Considering the findings of the 2016 Review and the review update, the CBR algorithm counted 17 studies and the trees were mentioned in 15 works. It is important to remember that a study about heterogeneous ensembles uses more than one group of methods. Also, it is interesting to notice the use of homogeneous ensembles in the formation of the heterogeneous ensemble.

Our findings show that the learning-based techniques remain the preference of the researchers. These results are consistent with the finding of Wen et al. (2012), in which they stated that the techniques using machine learning are the most widely used in the EEE literature. This could be explained because machine learning techniques tend to achieve better results than traditional techniques. It was also observed that the models of MLP, SVR and KNN are the most widely used algorithms for the evaluation of heterogeneous ensembles, whereas the homogeneous ensembles are usually composed of MLP, RT and CBR.

3.4. RQ3 - What are the different evaluation metrics, databases, matching rules, validation and statistical testing methods used in EEE?

In the answer to RQ03, we discuss the findings related to databases, metrics, combination rules, as well as the validation and statistical tests adopted in order to make the comparisons between EEE techniques.

As for the databases, PROMISE (Shirabad and Menzies, 2005) and ISBSG (Abran, 2015) repositories remain the main providers of software project data. Overall, we observed that 12 databases were mentioned in more than one study, as well as other databases that were mentioned once, with the majority of them being characterized as databases not available on the Internet. Table 10 summarizes the quantity of studies that made use of each dataset, comparing findings from the 2016 Review and the review update. Table 10 also presents the number of instances (size) and the percentage of use of each database. We would like to highlight the Cocomo81 and Desharnais databases, which have been used in over 50% of the studies between 2006 and 2020. These two databases are from the PROMISE repository (Shirabad and Menzies, 2005). In addition to being the most used, these datasets were more present in the relevant studies of the updated review. This fact is largely due to their popularity in previous studies and also to they relatively small size and the fact that they do not contain missing data.

The ISBSG repository was the third most used in just under 40% of the studies. However, it is important to highlight that it has been widely used quite recently. When compared with the 2106 Review, the findings of the review update show over 150% increase in the use of ISBSG, Albrecht and Kemerer databases. Albrecht was the third most prominent database from PROMISE, reaching over 33% of the studies. The ISBSG is a repository with a large amount of data, including information from different organizations, which makes it quite heterogeneous. The large amount of data, the heterogeneity and the possibility of partitioning the data into small sets may have led to the growth in the use of this repository. Meanwhile, the main reason for the relative increase in the use of Albrecht, Kemerer and Miyazaki datasets is because they were less used in previous years, probably due to the small number of instances of these datasets. However, in the new studies, the authors started to include these datasets believing that it is adequate to use them together with the leave-one-out validation method, which validates with greater robustness.

Moreover, the Tukutuku and China databases have become more prominent in the EEE scenario for the past 5 years. They had not yet been used in previous years. The China dataset was

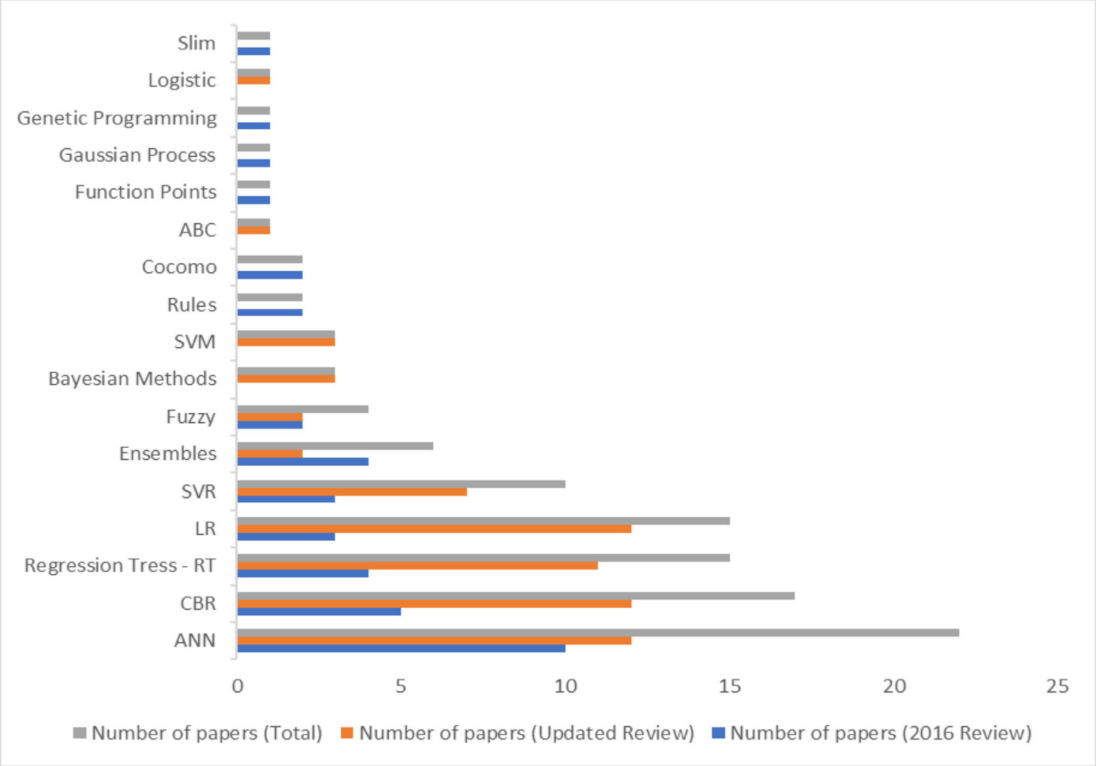


Fig. 8. Distribution of basic methods used in heterogeneous ensembles by studies.

Table 10
Number and percentage of studies by dataset.

Datasets	Size	2016 review	Percent	Updated Review	Percent	Total	Percent
Cocomo 81	63	11	45.8%	18	60.0%	29	53.7%
Desharnais	81	11	45.8%	16	53.3%	27	50.0%
ISBSG	10.769	6	25.0%	15	50.0%	21	38.9%
Albrecht	24	5	20.8%	13	43.3%	18	33.3%
Kemerer	15	4	16.7%	10	33.3%	14	25.9%
Miyazaki	48	4	16.7%	9	30.0%	13	24.1%
Nasa 93	60	6	25.0%	7	23.3%	13	24.1%
Maxwell	499	5	20.8%	7	23.3%	12	22.2%
China	93	–	–	9	30.0%	9	16.7%
Cocomonasa	62	7	29.2%	2	6.7%	9	16.7%
SDR	12	5	20.8%	2	6.7%	7	12.9%
Others	–	–	–	6	20.0%	6	11.1%
Tukutuku	53	–	–	3	10.0%	3	5.5%
Kitchenham	145	–	–	1	3.3%	1	1.9%
Apache; Jira; Spring; Jboos; MongoDB	–	–	–	1	3.3%	1	1.9%

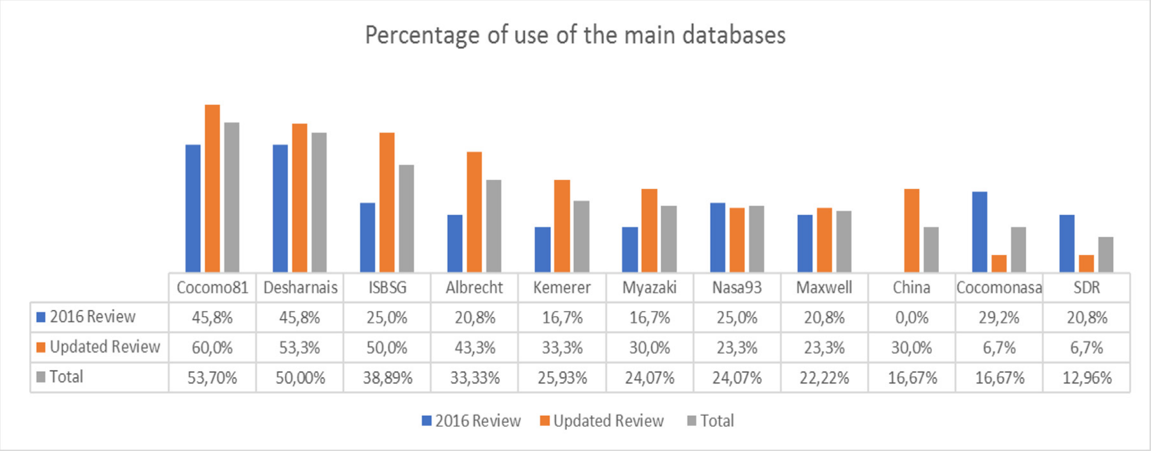
included later in the PROMISE repository and Tukutuku is hard to find publicly. The SDR databases are usually not considered in the studies probably because they have a small quantity of data. Fig. 9 presents the percentage of use of databases (over 10%) in the both reviews, in which it is easier to observe the increase, stability or disuse of the datasets.

The size presented in Table 10 is based on the original version of each dataset. The current version of ISBSG, according to (ISBSG, s.d.), showed more than 10.000 projects. However, this is not a public repository and each study has different sizes for this repository. The Apache, JBoss, JIRA, MongoDB, and Spring databases are part of projects that have been developed using the Scrum methodology, in which each iteration generates various work items. From these iterations, a total of 56.687 items were collected. The study (Gultekin and Kalipsiz, 2020) shows more details in terms of the use of these datasets.

As for the metrics, we identified 25 types of costs for the accuracy evaluation in EEE. Mean Magnitude Relative Error (MMRE - Eq. (1)) and Prediction (PRED - Eq. (2)) were the most commonly found metrics in the studies in the 2016 Review. In the review update, some metrics were frequent and showed an increase in relation to the 2016 review. It is likely that some of them tend to be more used in future studies. Metrics based on the MRE started to be used less, mainly due to the criticism presented in studies that show the bias of the MRE. Some criticisms about the MRE were shown in Shepperd and MacDonell (2012). Thus, the percentage of use of the MMRE, PRED and MdMRE decreased in the most recent studies. The MAE, probably due to its simplicity, and the SA possibly for presenting the results in percentage gains instead of a raw value, are the metrics that seem to be replacing those based on the MRE.

Table 11 shows the result of the quantity of studies that used each one of the metrics. It is noticeable the increase of the Mean Absolute Error (MAE - Eq. (3)) and the Standardized Accuracy (SA - Eq. (4)) metrics, as well as a decrease in the use of MMRE.

The absolute error (AE) is used for all cost functions presented above. In the SA metric, the MAE_{p0} is defined as the estimation error of a method, and the MAE_{pi} is the mean of a large number



	Cocomo81	Desharnais	ISBSG	Albrecht	Kemerer	Myazaki	Nasa93	Maxwell	China	Cocomonasa	SDR
■ 2016 Review	45,8%	45,8%	25,0%	20,8%	16,7%	16,7%	25,0%	20,8%	0,0%	29,2%	20,8%
■ Updated Review	60,0%	53,3%	50,0%	43,3%	33,3%	30,0%	23,3%	23,3%	30,0%	6,7%	6,7%
■ Total	53,70%	50,00%	38,89%	33,33%	25,93%	24,07%	24,07%	22,22%	16,67%	16,67%	12,96%

Fig. 9. Percentage of use of the main databases.

of random suppositions. SA suggests how good a method is when compared to random suppositions (Idri et al., 2016). The Median Magnitude Relative Error (MdMRE - Eq. (5)) is less used than the other metrics, but it reduces the failure of the MMRE. Finally, Mean Balanced Relative Error (MBRE - Eq. (6)) also minimizes MMRE bias by choosing the smallest of the values for the denominator of the equation. Table 11 also presents the amount and percentage of use of evaluation metrics in EEE.

$$MMRE = \frac{1}{n}\sum_{1}^{n}\frac{|actual\ value - estimated\ value|}{actual\ value} \quad (1)$$

$$PRED\ (x) = \frac{100}{n}\sum_{1}^{n}\frac{absolute\ error}{actual\ value} < x \rightarrow 1\ otherwise \rightarrow 0 \quad (2)$$

$$MAE = \frac{1}{n}\sum_{1}^{n}|actual\ value - estimated\ value| \quad (3)$$

$$SA = 1 - \frac{MAE\ pi}{MAE\ p0} \quad (4)$$

$$MdMRE = median\ (\sum_{1}^{n}\frac{|actual\ value - estimated\ value|}{actual\ value}) \quad (5)$$

$$MBRE = \frac{1}{n}\sum_{1}^{n}\frac{|actual\ value - estimated\ value|}{min\ (actual\ value\ ;\ estimated\ value)} \quad (6)$$

MAE is the most used metric as an input value in the statistical tests. All the other metrics are usually compared as numbers and graphics that are not derived from statistical tests. Other metrics should be emphasized because they are frequently used for effort estimation problems, such as MdMRE (Eq. (5)), MBRE (Eq. (6)), and others. Subsequently, we present the formula used by the most common metrics (higher than 20% in terms of use) in EEE problems. Fig. 10 shows the percentage of use of the most used metrics, considering the accumulated result of the studies from 2006 to 2020, more details can be found in Table 11.

The EEE schemes use different rules to combine the output of the base models. Considering the selected studies from the 2016 Review and the review update, we found 22 types of different rules that can be categorized as linear and nonlinear rules. A simple linear combination rule is the mean. It is the most common rule used in both homogeneous and heterogeneous ensembles. In general, EEE researchers have a preference for linear rules because they are simpler to use. The nonlinear combination rules are often more complex and meta-classifier models are used to join the outputs, for example, Stacking or the MLP neural networks. The simplicity of linear rules makes them more commonly used. Table 12 shows the quantity and the percentage of each combination rule for each combination rule, comparing the 2016 Review and the updated review.

Even though the median is less used than the mean, it tends to achieve better results in EEE problems (Cabral et al., 2017; Hosni et al., 2017; Hosni and Idri, 2017). Some works even suggested the use of the median among the combination rules analyzed. Nevertheless, from the evidence collected in both reviews, it is not possible to conclude that one combination rule is better than any other. Thus, we found it important to evaluate as many combinations as possible. It is likely that the increase in the percentage of use of the median and IRWM comes from the good results that these combination rules have obtained; Section 3.6 presents some studies that show these improvements.

We summarized the use of validation methods in Table 13. The leave-one-out method is the preference of EEE researchers, used in over 40% of the studies. This method is recommended for effort estimation due to the small size of the databases usually found in this area. Some works did not present the validation method, whereas others used more than one method in the same research. In that sense, the percentages shown in Table 13 were based on the amount of uses of each method and the total quantity of studies. We can see that the three validation methods have increased in terms of percentage of use (which was computed considering that some works use more than one of these validation methods). We believe that authors have started to give more importance to this criterion, which improves the quality of published articles. Hold-out grew less than the other methods, this fact is probably due to the size of the software effort estimation datasets. The more robust validation methods are more suitable for small databases, such as n-cross validation and especially leave-one- out. And they are now easily handled due to the increasing computational power over the years.

A total of 66% of the EEE studies used statistical tests for the purpose of comparing different techniques (individual/ compound). The nonparametric statistical tests are most common in the multiple comparisons. There is a preference for using the Wilcoxon and Friedman tests. Even though the nonparametric tests are widely used, the selection of the statistical tests depends on the data distribution and the number of groups that will be analyzed.

From the results presented in this section, we can conclude that:

- It is possible to use at least 5 databases in order to carry out tests for comparing methods;

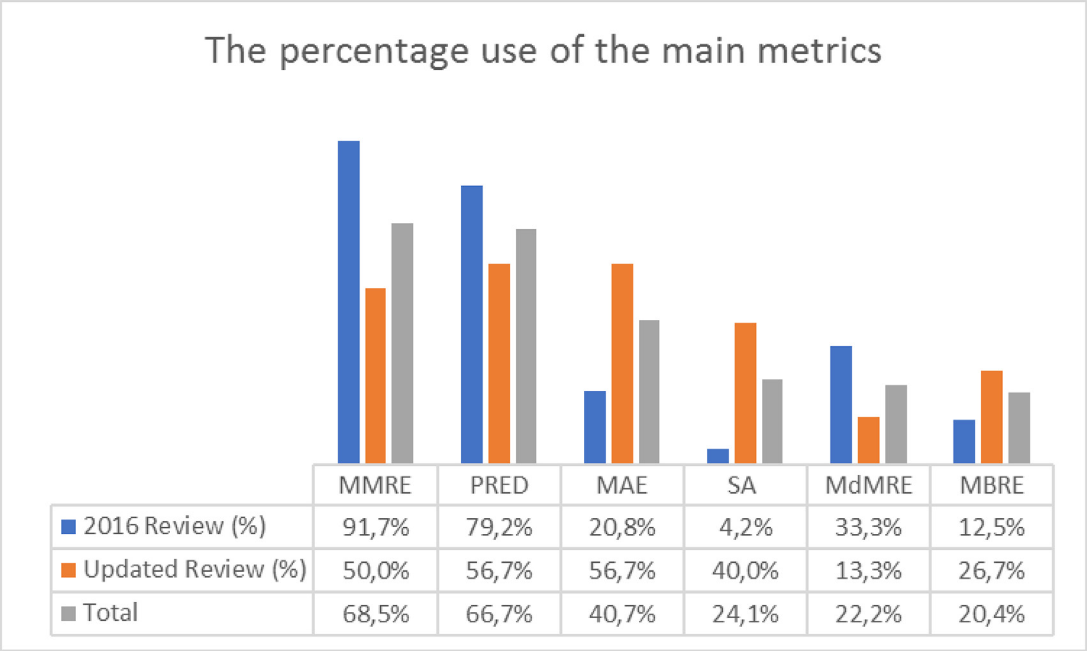


Fig. 10. Percentage of use of the most used metrics.

Table 11
Amount and percentage of use of the evaluation metrics.

Metric	2016 review	Percent	Updated review	Percent	Total	Percent
MMRE	22	91.7%	15	50.0%	37	68.5%
PRED	19	79.2%	17	56.7%	36	66.7%
MAE	5	20.8%	17	56.7%	22	40.7%
SA	1	4.2%	12	40.0%	13	24.1%
MdMRE	8	33.3%	4	13.3%	12	22.2%
MBRE	3	12.5%	8	26.7%	11	20.4%
Delta	1	4.2%	9	30.0%	10	18.5%
LSD	3	12.5%	7	23.3%	10	18.5%
MIBRE	3	12.5%	4	13.3%	7	13.0%
MSE	4	16.7%	1	3.3%	5	9.3%
RMSE	2	8.3%	2	6.7%	4	7.4%
MMER	3	12.5%	–	0.0%	3	5.6%
EF	2	8.3%	–	0.0%	2	3.7%
MdAE	1	4.2%	1	3.3%	2	3.7%
Accuracy	–	0.0%	1	3.3%	1	1.9%
BEM	–	0.0%	1	3.3%	1	1.9%
Corr	1	4.2%	–	0.0%	1	1.9%
MdMER	1	4.2%	–	0.0%	1	1.9%
MSRE	1	4.2%	–	0.0%	1	1.9%
NRMSE	1	4.2%	–	0.0%	1	1.9%
R	–	0.0%	1	3.3%	1	1.9%
RSD	–	0.0%	1	3.3%	1	1.9%
SD	1	4.2%	0	0.0%	1	1.9%
Smooth error	–	0.0%	1	3.3%	1	1.9%
STD pooled	1	4.2%	–	0.0%	1	1.9%

- The metric obtained from the absolute error is the one used as an input value of the statistical tests and the MAE use has increased in the EEE context.
- Although the SA metric is more uncommon, it has also shown an increase considering in the updated review;
- The linear rules are more common in EEE, especially the mean. Even though the median is less used than the mean, it tends to achieve better results in some EEE problems, such as in studies.
- The leave-one-out method is the most common and recommended validation method in EEE;
- The statistical tests for multiple comparisons more frequently use nonparametric techniques, especially the Wilcoxon and Friedman tests.

3.5. RQ4 - Do EEE techniques outperform the single models used in SDEE?

In this section, we synthesize the findings regarding the comparison of ensembles and single models. In the updated review, 27 studies, of the total of 30 selected studies, presented results of such comparison. Table 14 summarizes the results of these 27 studies.

Almost all studies (26/27) showed that ensembles outperformed base models, which confirms and increases the strength of the evidence found in the 2016 Review. Over 60% (17/27) of the studies stated that the ensembles outperformed the individual methods significantly, nine indicated that the improvements were not significant, and one stated that a base model was better

Table 12
Number of studies by combination rule.

Type	Combination rule	2016 review	Percent	Updated review	Percent	Total	Percent
Linear	Mean	18	75.0%	24	80%	42	77.8%
	Median	4	16.6%	10	33.3%	14	25.9%
	Inversed ranked weighted mean	2	8.3%	8	26.6%	10	18.5%
	Mean weighted	5	20.8%	6	20%	11	20.3%
	Median weighted	1	4.2%	2	6.6%	3	5.5%
	Geometric mean	1	4.2%	1	3.3%	2	3.7%
	Harmonic mean	1	4.2%	1	3.3%	2	3.7%
	Outperformance combination	1	4.2%	1	3.3%	2	3.7%
	Mode	–	–	2	6.6%	2	3.7%
	Equally weighted	1	4.2%	–	–	1	1.8%
	Weighted adjustment based on criterion	1	4.2%	–	–	1	1.8%
	IDWM	–	–	1	3.3%	1	1.8%
Non-linear	Meta classifier	–	–	3	10.0%	3	5.5%
	Support vector regression	2	8.3%	1	3.3%	3	5.5%
	Multilayer perceptron	2	8.3%	–	–	2	3.7%
	Multi linear regression	1	4.2%	1	3.3%	2	3.7%
	Fuzzy inference system using fuzzy c-means (FIS-FCM)	1	4.2%	–	–	1	1.8%
	Fuzzy inference system using subtractive clustering (FIS-SC)	1	4.2%	–	–	1	1.8%
	ANFIS-FCM	1	4.2%	–	–	1	1.8%
	ANFIS-SC	1	4.2%	–	–	1	1.8%
	RBF	1	4.2%	–	–	1	1.8%
	DENFIS	1	4.2%	–	–	1	1.8%

Table 13
Summary of the use of validation methods.

Method	2016 review	Percent	Updated review	Percent	Total	Percent
Leave one out	8	33.3%	15	50.0%	23	42.5%
n-cross validation	4	16.6%	10	33.3%	14	25.9%
Hold out	4	16.6%	8	26.6%	12	22.2%

than the ensemble evaluated. Considering the 17 papers where the ensembles improved the results significantly, 10 studies discussed homogeneous ensembles, four investigated both types of ensembles, and three used only heterogeneous ensembles.

Finally, the findings of the review update show that over 95% of the studies that compared individual models and ensembles recommended multiple estimation model systems and, among them, 63% showed an enhancement based on statistical tests.

3.6. RQ5 - Are there some EEE techniques that clearly outperform others?

In addition to making comparisons between ensembles and base models, some studies reported comparisons between different kinds of ensembles, combination rules, hyperparameter optimization, and attribute selection. In these evaluations, the rules, the models, and the types of combinations are factors that tend to influence the results.

In the 2016 Review, the authors observed that the ensembles based on SVR were more accurate than those based on DT and MLP, using comparisons considering the MMRE and PRED as cost measures. In the review update, we identified that 19 out of the 30 studies discussed comparisons of different kinds of ensembles. Our selected studies did not make comparisons between homogeneous and heterogeneous ensembles in order to identify the best type of ensemble, since the accuracy of the individual techniques contribute to the accuracy of the ensembles significantly. Subsequently, we will present an overview of the results from these studies.

However, it was still possible to extract some comparisons from the experiments carried out in the studies. The S8 study presented an exhaustive comparison between heterogeneous and homogeneous methods and demonstrated that the combination of heterogeneous methods using the median showed some superiority over the homogeneous methods that were evaluated. On the other hand, in the article (Twala and Verner, 2016) a slight superiority of Bagging and Boosting, which are types of homogeneous ensembles, over stacking, which is a popular algorithm for creating heterogeneous ensembles, was reported. Based on the results found in the 30 studies analyzed in the review update, we can say that as far as the types of ensemble are concerned, there is not sufficient evidence to state that there is some kind of superiority of homogeneous over heterogeneous ensembles.

In terms of the combination rules, the median and the IRWM (Inversed Ranked Weighted Mean) were found to be the most regular ones, especially in heterogeneous ensembles. The mean did not show the same accuracy when compared to the other rules. However, it is the most widely used combination and, in some cases, it shows similar accuracy to the other combinations, apart from being a simple rule concerning its implementation. In general, we observed that the weighted mean usually achieves better accuracy than arithmetic mean. There is not an overt superiority between the types of ensembles, instead there is a variation of winning techniques among the studies. The SVR was the most predominant individual method evaluated in the selected studies, probably because individual SVR obtains good estimates on SEE, but there is a lack of certainty concerning the superiority over the other methods. We present in Table 15 a summary of the selected studies.

3.7. RQ6 - Has dynamic model selection been used in the EEE literature?

In this section, we analyze the use of dynamic ensemble selection (DES) in software effort estimation projects. In the dynamic selection of models, a method or a set of methods are dynamically selected for each instance of the test (Ko et al., 2008). This research question is novel to the updated review.

When a set of models is previously aggregated through a combination rule to boost a value in the test set, there is a static ensemble, which uses the same set of base models and the same aggregation rules for all the elements in the test. Nonetheless,

Table 14
Comments regarding research question 04.

Study id	Comments
S1	The proposed method outperformed the best ANN base and traditional estimation methods.
S2	The estimations of the RF model provided lower MMRE values and higher PRED values than the MLP, RBFN, SGB and LLR estimation models.
S3	The empirical results showed that the proposed ensembles improved the accuracy of the estimations, taking into account the SA and PRED (25) on 2 out of 3 datasets compared to the individual models.
S4	The classic ensembles formed by analogy-based models outperformed the classic analogy solo techniques.
S5	Five ensembles created from Bagging, Boosting, Random Subspace, Stacking and Randomization outperformed the following individual models: ANN, DT, NBC, KNN and LgD.
S6	The study concluded that the performance of the base models is unstable, and some model combinations can achieve stability and a better performance compared to the individual models.
S7	The study presented various ABE schemes. Nonetheless, none of the 7 methods used achieved better results within all contexts considered.
S8	This work presented a broad comparison between various types of ensembles and individual models. The dynamic ensembles also outperformed the individual models.
S10	The heterogeneous ensembles were built using three combination rules over seven datasets. The ensembles achieve higher accuracy than their single techniques whatever the optimization technique used; in general, heterogeneous ensembles provided more accurate estimation.
S11	The study made a comparison on 7 databases. In 6 of them, the best ensemble outperformed the best solo model.
S12	The study presented an ensemble constructed by SVM, MLP and GLM, considering the ISBSG databases. The results showed that SVM outperformed the ensemble formed by the mean of the individual estimations. This study may have failed, because it did not present results using the median or IRWM.
S13	The results presented indicated that an ensemble of Optimal Trees outperforms the base regression trees significantly.
S14	The study showed that the use of ensemble is a good alternative approach when the base models are not consistent in different datasets, as well as when the models behave in distinct ways. However, significant differences were not spotted in the results obtained.
S15	The study evaluated the impact of 3 filters (attribute selection) on the predictive capacity of single and ensemble techniques. A statistical test showed that there is no statistical difference between the ensemble and single techniques except in one dataset, however, the evaluations suggested that fuzzy analogy ensembles underperformed their individual members when feature selection was used.
S17	The study compared an RF model with MLP and SVM on a personal database. The results showed that the RF was found to have a better performance than the individual models.
S18	The study proposed an ensemble model (SEET) constructed from the integration of SA-based models using MLP, LSSVR and ELM. The SEET model was found to be better, in most of the cases, than the individual models.
S19	The study constructed a RF and compared it with RTs. The authors concluded that none of the simple methods showed a superiority over the others in any situation, but the results showed that the RF outperformed RTs within all the evaluation criteria.
S20	The study presented comparisons between individual methods and a multiple combination method proposed. An ensemble with generated weights that outperformed the individual methods was presented, taking into consideration a personal database and the results obtained from statistical tests.
S21	The study showed the superiority of various types of ensembles over the base method. The article used an MLP as an individual estimation strategy and combined the method in different ways. All the ensemble outperformed the individual model based on the results presented.
S22	The study investigated the design of a Random Forest model. The model was compared with a regression tree model on 3 datasets and 3 evaluation metrics. The Random Forest model outperformed the models based on regression trees considering all the evaluation criteria.
S23	The study made a robust comparison with 14 widely used algorithms in the data science community. The findings demonstrated that the superiority of ensemble-based algorithms embrace EEE studies, as well as other domains.
S24	The study showed the superiority of an ensemble constructed from the bootstrapping technique using the RVM algorithm in relation to individual methods, such as KNN, SVR and MLP.
S25	The study presented the comparison of a regression tree Boosting with results obtained from the literature through other methods. The authors believe that the methodology is promising, because it shows better results than the ones presented in the articles.
S26	The study constructed a heterogeneous ensemble based on majority voting and characteristic extraction. The performance of the proposed model is encouraging when compared to the individual model.
S27	The study conducted an experiment across a personal database considering different stacking and bagging ensembles. The comparisons determined that the proposed ensemble model can increase the efficiency, reduce the error rate and improve the predictive accuracy in relation to the individual models.
S28	The study proposed a hybrid model of ABC with ABE using weights that are calculated in the training phase. The proposed model is more accurate than the different individual estimation models using ABE.
S30	The study investigated the results obtained from individual models and ensemble models. Boosting and Random Forest models outperformed SVR and MLP in most of the datasets evaluated.

when a different set of models is selected for each instance of the test set, then there is a dynamic selection. As for the latter, the set of base models for each instance of the test is determined when the estimation is taking place. The dynamic model makes a selection of the potentially best base models for a specific instance. This selection is usually based on the set of attributes of the evaluated instance.

Only 2 studies mentioned the concept of dynamic selection. Twala and Verner (2016), presented experiments where the ensembles, which were created from the Bagging and Boosting algorithm, achieved better results than those of a simple dynamic selection. The simple selection dynamically chooses a single base model to carry out the estimation. However, the authors have not mentioned in detail how this dynamic method was created.

Cabral et al. (2017) proposed a new dynamic selection method and applied it to estimation databases of SEE. The proposed model has three variations: a simple dynamic selection (DS) and two dynamic ensemble selections (DES). The first DES model used the median, while the latter used the mean as a combination rule. The experiments reported in the paper showed that the proposed models were superior to all the static ensembles and individual methods. We believe that the DS and DES process applied to SEE

Table 15
Comments regarding research question 05.

Id	Comments	Type of comparisons
S3	The study evaluated 3 rules. The IRWM was the most suitable combination in 2 out of 3 databases, whereas none of the combination rules won in the other dataset.	Rules
S4	The study evaluated ensembles based on analogy. The authors concluded that the results on the databases greatly vary in terms of the combination rules and the quantity of base models. Furthermore, significant differences were not found between the 2 combination rules used (Classic and Fuzzy).	Rules
S5	The study evaluated the performance of different schemes. Bagging achieved the best accuracy in comparison with Boosting, Stacking, and Randomization.	Ensemble types
S6	The study evaluated 9 combination rules and concluded that 6 of them were not significantly different (mean, median, median weighted, wcr combination, mean weighted and inverse ranked weighted mean); however, 3 other combination rules (harmonic mean, outperformance combination and geometric mean) were worse than the others.	Rules
S7	The study made various comparisons, and the authors concluded that the WMedian was the best rule in relation to the rules investigated.	Rules
S8	The study evaluated static and dynamic ensembles. The static evaluation showed that the median used in the heterogeneous ensemble achieved a better result when compared to the other ensembles. Moreover, a comparison with homogeneous ensembles showed that Bagging obtained a better result than the Boosting in a general way, but there were no significant differences. This study also revealed that the dynamic ensembles selection outperformed the static ensembles.	Rules/Ensemble types
S9	The study evaluated ensembles with and without the attribute selection. The authors concluded that all the attributes, in the majority of the datasets, were important in order to generate a predictive and accurate model. Besides that, the study evaluated three types of combinations: average, median and inverse ranked weighted mean. The median led the ensembles to achieve better results from a general perspective.	Rules/Attribute selection
S10	The study evaluated ensembles with the use of methods of parameter optimization of algorithms. The results showed that Grade Search (GS) and Particle Swarm Optimization (PSO) achieved better results than the standard configuration of the Weka library. Furthermore, the latter and the former caused, in general terms, the same impact in relation to the accuracy achieved.	Optimization
S11	The study presented a variation in the results among the databases regarding the comparison among the homogeneous ensembles that used median and the mean. However, the result showed a superiority of the median on 5 out of 7 databases, whereas the mean stood out on 2 datasets.	Rules
S13	The study showed that the RF was outperformed by an ensemble of Optimal Trees, which leads one to believe this is a promising technique in SEE.	Ensemble types
S15	The study analyzed ensembles with and without the attribute selection, plus some combination rules. The results revealed that there is not the best ensemble for all datasets. However, the authors concluded that FA ensembles without feature selection outperformed FA ensembles with filter feature selection in the most of cases analyzed. As for combination rules, the IR achieved the best accuracies on most datasets, followed by the median and the mean.	Rules/Attribute selection
S16	The study constructed 3 types of ensembles: ENF that used N filters of characteristic selection, E0F that did not select any attributes, and E1F that only used a single filter of attribute selection. Fifteen different combinations using KNN, MLP, SVR and DT were evaluated. The experiments indicated that there is not the best ensemble, however E0F was found to have the highest value of SA on 4 datasets, ENF on 2 datasets, and E1F did not reach first place on any of the datasets. As for the combination rules, the IRWM and the median generated the best SA values.	Rules/Attribute selection
S18	The study presented a comparison between two types of ensembles. One was constructed from a suggested model, and the other was created from the IRWM combination rule. The study results carried out an analysis on 6 databases showed that the model of the suggested ensemble won on 66% of the databases analyzed. These findings indicated that the results can be improved if the strategies for constructing the ensemble are not restricted to simple rules of linear combinations.	Rules/Ensemble types
S20	The study (S20) presented a new ensemble using a combination of weights for each individual method. The method used 7 prediction algorithms for estimating the effort of stories used in agile software development. The suggested ensemble outperformed the estimation means, Random Forest, Bagging and Boosting.	Ensemble types
S21	The study compared 5 MLP schemes, more precisely a simple MLP, an ensemble Ridge-MLP, an ensemble Lasso-MLP, a Bagging MLP and a Boosting MLP. The results obtained in this study demonstrated that the R2 presented by the Boosting MLP was superior to the other models.	Ensemble types
S23	The study drew a robust comparison with various algorithms. Bagging and Random Forest outperformed Boosting on most of the databases. Stacking also received a great evaluation and outperformed many models.	Ensemble types
S24	The study showed the superiority of an ensemble scheme using bootstrapping (SynB-RVM) in relation to the other ensemble schemes, such as Bagging-RVM, Bagging-SVR.	Ensemble types
S29	The study compared different Stacking schemes. This work investigated the use of hyperparameters using PSO and GA. An optimization of hyperparameters of base models and the meta-learning algorithm is achieved in 2 approaches. Stacking-PSO showed a slightly superior accuracy when compared to Stacking-GA, as well as surpassed the results obtained by the Stacking without the optimization of hyperparameters.	Ensemble types/Optimization
S30	The study investigated a comparison with various models. A Random Forest model and a Boosting model were analyzed on 5 databases. The Boosting model outperformed the RF one considering all the comparisons.	Ensemble types

should be further investigated, since these are techniques that are consolidating and being widely used in the machine learning literature (Ko et al., 2008; Hou et al., 2020; Muhammed Niyas and Thiyagarajan, 2021; Cruz et al., 2018, 2020). A challenge encountered through this study was the search of papers that investigate the application of dynamic selection methods in EEE. A very recent study, published in 2021 (Cabral and Oliveira, 2021), addresses in more detail the importance of dynamic selection and the gains made in software effort estimates.

3.8. Summary of the findings

In this Section, we summarize the main findings of the review update:

- The proportion of studies that used homogeneous ensembles decreased when compared to the heterogeneous ensembles in the past 5 years (from 70% to 60%). Nevertheless, 17 of the studies used homogeneous ensembles, whereas 16 investigated heterogeneous ensembles. Moreover, we must highlight that some researchers discussed the two types of ensembles in the same study.
- Neural networks were the most commonly used techniques in homogeneous and heterogeneous ensembles with an emphasis on MLP networks.
- The mean was the combination rule adopted in most of the studies, followed by the median and the IRWM. However, median and IRWM achieved best results than the mean in general.
- Cocomo81 and Desharnais were the most commonly used databases in the EEE researches; however, Albrecht, ISBSG and Kemerer use has increased in the past few years. Tukutuku, China and some proprietary and open databases were not discussed in the previous work (Idri et al., 2016), even though they have been growing in the past few years and tend to continue to be used. The use of Cocomonasa and SDR has decreased.
- MMRE, PRED and MdMRE are still widely used, however, there has been an overt increase in the use of MAE and SA (Standard Accuracy). With respect to MAE, the metric has been widely used for the past few years, and the SA had a growth in percentage. These changes on the used of metrics have been influenced by the bias found in MMRE, which will be later discussed.
- Leave-one-out was by far the most commonly used validation method. This probably occurred because of the relatively small size of the databases.
- *Wilcoxon* and the Friedman test are common statistical tests used in the studies that presented the results through statistical experiments. The data analysis tends not to follow a regular distribution, which leads the comparisons to be made with non parametric tests in EEE researches.
- EEE improves the accuracy of the individual methods. Over 95% of the relevant studies that this work gathered presented enhancement in the use of ensembles in software effort estimation, only one study failed to show improvements.
- The median and the IRWM have taken a small advantage over ensembles created using other combinations, but we cannot state that they are the best way to combine individual methods for EEE, since there is no combination rule that outperforms the rest.
- Dynamic selection is still little used in EEE for software effort estimation, however, some results have shown that this kind of approach can lead to an improvement in the results compared to static ensembles.

Table 16 presents a summary of the main findings of the 2016 Review of Idri et al. (2016) together with the respective challenges or gaps identified.

Next, we present some comments from our updated review on the challenges presented in Table 16:

- The studies presented in our review do not support an identification of the best kind of ensemble for a given dataset nor in the general context. However, the results of study S8 show that heterogeneous ensembles outperformed both Bagging and Boosting ensembles (using the same individual learning methods). In addition, study S5 showed a superior performance of Bagging and Boosting compared to Stacking.
- MMRE is still the more commonly used metric in EEE studies, yet it is being replaced by MAE and SA due to its bias (Shepperd and MacDonell, 2012). MRE based metrics may favor minimum or maximum estimates. MAE is commonly used in studies that also employ statistical tests; we believe that it will become the standard metric in EEE.
- The mean is frequently used as the combination rule in homogeneous ensembles; study S18 used a weighted mean. The IRWM and the median have provided the best results in most of the studies using heterogeneous ensembles. Studies S7, S8, S9 and S11 achieved good results using the median, whereas in the S3 and S16 studies, IRWM was the best combination rule.
- More than 60% of the studies used statistical tests to confirm the superior performance of ensembles compared to individual methods. Yet, most of these studies did not present sufficient details about how these statistical tests were performed.
- To determine what are the EEE techniques with relatively higher performance in terms of estimation accuracy is not an easy task. There is not a clear advantage in favor of any of the techniques.
- Homogeneous ensembles generate models with different number of individual learners depending on the training hyperparameters. The studies do not present a conclusion on the ideal number of models in homogeneous ensembles. In the studies with heterogeneous ensembles, the ensembles had between 3 to 5 individual models. In general, SVR is the best performing algorithm in heterogeneous ensembles.

To better understand the differences between our findings and those reported by the 2016 Review by Idri et al. (2016), we compare and aggregate our findings with those of Idri et al. (2016) and present them in Table 17.

4. Limitations and mitigations of the review update

A common limitation in systematic reviews is related to search coverage. Ensuring that we find all studies that discuss the research questions is the greatest challenge. These are also the main limitations of this review. We tried to minimize this limitation by using well-established guidelines for our research protocol. Further, we used the same search string from the 2016 Review, that had already been validated. Three bibliographic sources were used for the purpose of collecting the highest quantity of studies. Finally, a snowballing search was carried out from the result of the automatic search. We believe that it improved the coverage of the review, since part of the relevant articles were selected from the manual search. It is possible that some papers had not been selected because they were not indexed in the search engines at the time that we did the search (March-2021). In addition, this research only considered studies published up to 2020.

The references and studies that mentioned the relevant articles were analyzed so that new articles that could answer the

Table 16
Summary of Idri et al. main findings and challenges.

Research question	Purpose	Main findings	Challenges or gaps identified
RQ1 What are the most frequently used EEE techniques?	To identify the EEE techniques used in literature as well as the learners that have been employed by researchers to build ensembles.	– Homogeneous and heterogeneous ensembles were analyzed. – Homogeneous ensembles have been more used than heterogeneous ensembles. – Artificial Neural Networks and Decision Trees are the most common learners	– To define the best kind of ensemble for EEE. – To define the best kind of ensemble for a given dataset.
RQ2 What is the overall estimation accuracy of EEE techniques?	To discuss the accuracy of EEE techniques in terms of four aspects: accuracy criteria, accuracy values, data sets, and evaluation methods.	– Accuracy of EEE techniques were studied in terms of four aspects: accuracy criteria, accuracy values, data sets, and evaluation methods. – Cocomo81 and Desharnais were the most frequently used data sets. – The dominant validation methods were holdout, n-fold cross validation and leave-one-out cross validation. – 16 evaluation metrics were found, with MMRE (Mean Magnitude of Relative Error), Pred(25) (Percentage of Predictions that are within 25% of the real value), and MdMRE (Median Magnitude of Relative Error) being the most used metrics. – Regarding accuracy: (i) homogeneous ensembles based on DT are the most accurate than others ensembles considering MMRE and PRED as an evaluation metrics, whereas (ii) heterogeneous ensembles are more accurate if MMdRE is used.	– To identify if new performance metrics, other than those based on MRE, are more appropriate for EEE. – To identify metrics used together with statistical tests.
RQ3 What are the different rules used to combine estimation models?	To identify the different rules used when combining single models through EEE techniques.	– 18 rules were found and categorized in two types: linear and non-linear. – Linear rules were the most used ones. – Homogeneous ensembles using the mean formula as the combiner are the most accurate ones. – Heterogeneous ensembles using median as the combination rule are the most accurate ones.	– Investigate new combination rules for homogeneous ensembles. – Investigate whether any combination rule outperforms the median in heterogeneous ensembles.
RQ4 Do EEE techniques outperform the single models used in SDEE?	To determine whether EEE techniques outperform conventional single techniques in terms of estimation accuracy.	– Homogeneous ensemble techniques are more accurate than single models. – Heterogeneous ensembles outperformed the single models in 75% of the evaluations based on the MMRE criterion, and they outperformed the single models in 80% of the evaluations based on the Pred criterion.	– To analyze the superiority of ensembles considering statistical tests and metrics not based on MRE (to avoid bias) – To have more information on the statistical tests performed to verify if the chosen test was adequate to the problem.
RQ5 Are there some EEE techniques that clearly outperform others?	To determine the EEE techniques with relatively high performance in terms of estimation accuracy.	– Ensembles based on Support Vector Regression models seem to be more accurate than those based on Decision Trees and Multilayer Perceptron	– To identify the best combination rule or the ideal number of base models.
RQ6 How to construct ensembles?	To identify the methodologies used to select the base models used when constructing an ensemble.		– Number of base models that will produce higher accuracy is an issue that has not been given attention by researchers and it can affect the accuracy. – Choosing the base models of an ensemble is a great challenge and must be made carefully, that is, the "right" models for a given context is a critical task.

research questions could be found. Then, criteria based on eight quality assessment questions were applied to determine the quality of the relevant studies. The gathering of relevant articles was carried out by 2 researchers: the first and second authors of this work. The third researcher (expert in systematic literature reviews) revised the process at the end. The first author carried out

Table 17
Comparing our findings with Idri et al.

Research question (our review)	Idri et al.	Our findings	Aggregation of results
RQ1 What are the most frequently used EEE techniques (homogeneous/heterogeneous)? To identify the EEE techniques used in literature. RQ2 What are the most frequently used individual techniques in the composition of EEE? To identify the learners that have been employed by researchers to build ensembles.	– Homogeneous ensembles have been more used than heterogeneous ensembles. – Artificial Neural Networks and Decision Trees are the most common learners.	– Homogeneous ensembles are superior in terms of the total number of studies, but the use of heterogeneous ensembles has grown more than the use of homogeneous ones. – Artificial Neural Networks and Regression Trees keep being the most commonly used base methods in homogeneous ensembles, while in heterogeneous ensembles ANN, CBR, RT, LR and SVR are the most common.	– Findings related to the use of homogeneous ensembles and ANN and RT as the most common learners are consistent between the two studies. – In the review updated, the use of heterogeneous ensembles has grown more than the homogeneous ones. Also, other learners were found in the heterogeneous ensembles.
RQ3 What are the different evaluation metrics, databases, matching rules, validation and statistical testing methods used in EEE?	– Linear rules were the most used ones. – Homogeneous ensembles using the mean formula as the combiner are the most accurate ones. – Heterogeneous ensembles using median as the combination rule are the most accurate ones.	– The findings of Idri et al. were confirmed in this study. – New datasets appeared while others became little used. ISBSG and PROMISE data remain the main repositories. – MAE and SA tend to replace MRE-based metrics. – Leave-one-out remains the most common validation method, probably due to the size of the databases. – The use of statistical tests is still not so common in studies, but they have become more used.	– The review update increased the strength of the evidence for RQ3. – Results have been refined and extended in the review update.
RQ4 Do EEE techniques outperform the single models used in SDEE?	– Homogeneous ensemble techniques are more accurate than single models. – Heterogeneous ensembles outperformed the single models in 75% of the evaluations based on the MMRE criterion, and they outperformed the single models in 80% of the evaluations based on the Pred criterion.	– The results presented in this study confirm what was shown in the previous one. 63% of the studies showed an enhancement based on statistical tests and over 95% of the studies recommended ensemble systems.	– The review update increased the strength of the evidence for RQ4.
RQ5 Are there some EEE techniques that clearly outperform others, regarding type, rules and composition?	– Ensembles based on Support Vector Regression models seem to be more accurate than those based on Decision Trees and Multilayer Perceptron	– SVR has shown good results individually, but it was not possible to confirm the finding by Idri et al. Nor is it possible to say which type of ensemble is better, homogeneous or heterogeneous. However, some combination rules, such as the median and the IRWM, have shown better results.	– The review update did not increase the strength of the evidence for RQ5. The aggregation of the results indicated that further primary studies are necessary.
RQ6 Has dynamic model selection been used in the EEE literature?	Not addressed by Idri et al.	– Only two studies investigated dynamic selection of models in ESS, in one of them the results were very satisfactory and surpassed the static ensembles.	– The review update extended the 2016 Review with RQ6.

Table 18
Preliminary data extraction fields.

Fields	Description	Research question
Id	Study identification.	Descriptive
Year	Year of publication.	Descriptive
Authors	Names and surnames of the study authors.	Descriptive
Source	Type of publication: Conference or Journal.	Descriptive
DOI	Global identification of the selected study.	Reference
URL	Electronic location: Address in web format.	Reference
Title	Study title.	Reference
Homogeneous ensemble	Identify whether the study used a homogeneous ensemble, and which techniques (Bagging, Boosting, etc.) were addressed.	RQ1, RQ2
Heterogeneous ensembles	Identify whether the study used a heterogeneous ensemble, and with which individual techniques.	RQ1, RQ2
Base methods	Identify the basic methods used in the ensemble and in the experiments.	RQ2
Datasets	Identify the datasets used in the experiments.	RQ3
Estimation criteria	Identify the cost metrics used in the experiments.	RQ3
Validation methods	Identify the validation methods used to evaluate machine learning models.	RQ3
Matching rules	Identify the combination rules used to compose the aggregate result of the ensemble.	RQ3
Simple assessments	Identify if there were any comparisons between individual methods and ensembles, and record the winners.	RQ4
Composite evaluations	Identify if there were existing comparisons between the different types of ensembles. Homogeneous × Homogeneous/Homogeneous × Heterogeneous/Heterogeneous × Heterogeneous.	RQ5
Dynamic methods	Identify whether there were dynamic selection methods in the research.	RQ6
Dynamic comparisons	Identify there were comparisons between ensembles and dynamic methods.	RQ6
Statistical tests	Identify whether statistical tests were used in the experiments.	RQ4, RQ5, RQ6

Table 19
Expanded data extraction.

Fields	Means	Type
Id; Title; Year.		Text
Author 1; Author 2; Author3; Author 4; Author 5.		Text
Source (see Updated review in Fig. 3).	Conference/Journal	Text
Bagging; Boosting; Random Subspace; Random Forest; Others (see Fig. 6).	Homogeneousensemble	Boolean by ensemble
Stacking	Heterogeneousensemble	Boolean
ABC; ABE; ANN; BN; CART; CBR; DT; ELM; Fuzzy; GLM; KNN; LaR; LgD; LMS; M5P; MLP; MLR; NB; OLSR; RF; RBF; RR; RT; RVM; SR; SVM; SVR (see Updated review in Figs. 7 and 8).	Base methods	Boolean by algorithm
Heterogeneous	If there was use of heterogeneous ensemble?	Boolean
Albrecht; Apache; Cocomo81; China; Desharnais; ISBSG; Jira; JBoss; Kemerer; Kitchenham; Maxwell; Miyazaki; MongoDB; Nasa1; Nasa2; Spring; Tukutuku and others (see Updated review in Table 10).	Datasets	Boolean by dataset
Accuracy (hit rate); Delta; LSD; MAE; MBRE; MMRE; MdMRE; MIBRE; MSE; MdAE; MdBRE; MdIBRE; PRED; R; RMSE; RSD; SA; Smooth Error (see Updated review in Table 11).	Metric	Boolean by metric
Leave-one-out; Cross-validation; Hold-out (see Updated review in Table 13).	Validation methods	Boolean by method
Mean; Median; Mode; IRWM; Median W; Average W; Geometric Mean; Harmonic Mean; Outperforming Combination; Non linear (see Updated review in Table 12).	Matching rules	Boolean by rule
Is the proposed model hybrid?	If a hybrid model was used	Boolean
Did the ensemble win?	If the proposed ensemble won	Boolean
Did the study perform statistical tests?	If any statistical method was used in the experiment	Boolean

the initial search and pre-selected the papers, the second author supervised the process, reviewing the pre-selected articles, and participating in data extraction, as discussed in Section 2.

5. Concluding remarks

This systematic review updated the research in EEE presented by Idri et al. (2016). EEE techniques were examined from six perspectives: (i) frequency of use of homogeneous and heterogeneous ensembles; (ii) individual models adopted to create ensembles; (iii) metrics, databases, combination rules, validation and statistical tests used in EEE studies; (iv) the superiority of EEE techniques over individual models; (v) superiority among EEE techniques; and (vi) use of dynamic ensembles in SEE. An automatic search was carried out in three bibliographic sources in order to select the studies that could answer our research question. Snowball search was also carried to increase coverage. The search considered works developed between 2016 to 2020 and a total of 30 relevant studies were selected. Our main findings are summarized below.

- Which types of ensembles (homogeneous/heterogeneous) are the most widely used in effort estimation?

The 2016 Review showed that the homogeneous ensembles were the most widely used. They were supported by over 70%

(17/24) of the selected studies, whereas the heterogeneous ones were present in just under 38% (9/24). Some articles studied both homogeneous and heterogeneous ensembles. The review update found that the use of heterogeneous ensembles has increased more than the use of homogeneous ensembles.

- What is the frequency of use of individual techniques in the construction of the EEE?

The 2016 Review identified 16 estimate models used to construct both types of ensembles. The machine learning models were preferred, with ANN and DT being two single models most frequently used; in particular, MLP is the most common form of ANN. In the updated review, studies still prefer learning-based techniques. The neural networks are the most widely used base methods in the homogeneous and heterogeneous ensembles, considering all studies presented in the 2016 Review and the review update. The MLP was the most common type of ANN. Regression trees and the similarity-based models were found to follow learning-based models as the next most widely used techniques in homogeneous and heterogeneous ensembles, respectively. The trees are the second most used technique in homogeneous ensembles and the similarity-based models are the second most used in heterogeneous ensembles.

- What are the different evaluation metrics, databases, combination rules, validation methods and statistical tests used in EEE?

The 2016 Review identified 16 identified that Cocomo81 and Desharnais datasets as the most used and this trend was confirmed in the review update. The sources from PROMISE and ISBSG remain the main providers of software project data. Cocomo81 and Desharnais from PROMISE, and the data from ISBSG, were found to be prominent. Furthermore, a significant increase in the use of the China and ISBSG databases was observed in the review update.

As for the metrics, MMRE, PRED and MdMRE were the most predominant in the studies from the 2016 Review. In the review update, new metrics, such as MAE, SA and Delta, stood out and showed a tendency to be more widely used in recent studies. This may be because MMRE is considered a biased metric (Shepperd and MacDonell, 2012).

As in the 2016 Review, the mean is the most common combination rule, especially when it comes to homogeneous ensembles. The leave-one-out method is the most frequently used validation method regarding effort estimation papers, repeating what was found in the 2016 Review. However, depending on the size of the databases, the cross-validation (k-fold = 5 or 10) or the hold-out method were also applied. The nonparametric statistical tests are more common in multiple comparisons since the data are usually not normalized. The value of the MAE or AR is the most commonly used metrics as an input either to parametric or nonparametric statistical tests.

- Have EEE or adaptations of compound models outperform the individual models?

In the 2016 Review, the authors found empirical evidence that, in the majority of cases, EEE techniques were more accurate than single models. In the review update, 96% (26/27) of the studies that made comparisons between individual techniques and ensembles suggested that the use of combined techniques improved performance. Over 60% (17/27) of these studies showed superiority of EEE through statistical tests.

- Which configurations in EEE (regarding the type, the rules, and the construction) have outperformed the others?

According to the findings of the 2016 Review, it was not possible to state any EEE technique as being the most accurate. Consistently, in the review update we still did not have enough studies to attest some superiority of one of the types of ensembles (homogeneous or heterogeneous). In terms of the combination rules, the median and the IRWM were found to be the most regular linear rules, the SVR was seen to be the most prominent individual method in the construction of the heterogeneous ensembles, and MLP and RT were found to be the best algorithms in the construction of the homogeneous ensembles.

- Has the dynamic selection of the models been used in the EEE literature?

Only two studies in the review update mentioned the concept of dynamic selection. Moreover, only one of them showed comparative results between dynamic and static ensemble methods, showing the advantages to the dynamic ensembles, which can be seen as an opportunity for new studies to investigate dynamic selection strategies.

For future works, we can see an opportunity for new studies that use different datasets of software effort estimation since most of the studies so far perform their experiments in the same datasets. The PROMISE and ISBSG repositories are most commonly used. However, some studies used private datasets, which is good to achieve new conclusions about the software estimate, but when these datasets are not available, the replication is not possible, and this hampers the advancement of studies in software effort estimation. Another approach that can been investigated to reach more accurate results is the use of alternative strategies capable of combining the regressors in different ways, e.g, the use of Dynamic Ensemble Selection (DES) methods. Few studies have addressed the use of DES in EEE. However, DES is a strategy that has obtained good results in prediction problems in general, and it is still little used in a context in which the dependent variable is a real value. In this sense, future studies could follow in these directions in order to find out new samples and strategies for the estimation of software effort.

CRediT authorship contribution statement

José Thiago H. de A. Cabral: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Adriano L.I. Oliveira:** Validation, Writing – review & editing, Supervision. **Fabio Q.B. da Silva:** Validation, Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

Some of the data used can be share. But we also used data that cannot be shared.

Acknowledgments

Fabio Q. B. da Silva receives a research grant from CNPq, Brazil #303738/2020-0. Adriano L. I. Oliveira receives a research grant from CNPq, Brazil # 314499/2021-0.

Table 20
Selected studies.

Study id	Reference
S1	H. Ibrahim and B. H. Far, "Clustering and Artificial Neural Network Ensembles Based Effort Estimation," em Published in SEKE 2016, 2016.
S2	S. M. Satapathy, B. P. Acharya and S. K. Rath, "Early stage software effort estimation using random forest technique based on use case points," *IET Software,* vol. 10, n° 1, pp. 10–17, 2016.
S3	M. Hosni, A. Idri, A. B. Nassif and A. Abrain, "Heterogeneous Ensembles for Software Development Effort Estimation," em *3rd International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Dubai, 2016.
S4	A. Idri, M. Hosni and A. Abran, "Improved estimation of software development effort using Classical and Fuzzy Analogy ensembles," *Applied Soft Computing,* vol. 49, pp. 99–1019, 2016.
S5	B. Twala and J. Verner, "Toward Accurate Software Effort Prediction Using Multiple Classifier Systems," *Computational Intelligence and Quantitative Software Engineering,* vol. 617, pp. 135–151, 2016.
S6	S. Amasaki, "A Comparative Study on Linear Combination Rules for Ensemble Effort Estimation," em *43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Vienna, 2017.
S7	A. K. Bardsiri, "A new combinatorial framework for software services development effort estimation," *International Journal of Computers and Applications,* pp. 14–24, 2017.
S8	J. T. Cabral, A. Oliveira, J. Nóbrega and R. Araújo, "Heterogeneous Ensemble Dynamic Selection for Software Development Effort Estimation," em *IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, Boston, MA, 2017.
S9	M. Hosni, A. Abran and A. Idri, "Investigating Heterogeneous Ensembles with Filter Feature Selection for Software Effort Estimation," em *12th International Conference on Software Process and Product Measurement*, Gothenburg, Sweden, 2017.
S10	M. Hosni, A. Idri, A. Abran and A. B. Nassif, "On the value of parameter tuning in heterogeneous ensembles effort estimation," *Soft Computing,* vol. 22, n° 18, pp. 5977–6010, 2017.
S11	M. Hosni and A. Idri, "Software effort estimation using classical analogy ensembles based on random subspace," em *Proceedings of the Symposium on Applied Computing*, Marrakech, Morocco, 2017.
S12	P. Pospieszny, B. Czarnacka-Chrobot and A. Kobylinski, "An effective approach for software project effort and duration estimation with machine learning algorithms," *The Journal of Systems and Software,* vol. 137, pp. 184–196, 2018.
S13	A. Zakrani, M. Hicham and N. Abdelwahed, "An Ensemble of Optimal Trees," em *International Conference on Advanced Information Technology, Services and Systems*, 2018.
S14	M. Azzeh, S. Banitaan, A. B. Nassif and C. L. Martín, "Ensemble of Learning Project Productivity in Software Effort Based on Use Case Points," em *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, 2018.
S15	M. Hosni, A. Idri and A. Abran, "Evaluating filter fuzzy analogy homogenous ensembles for software development effort estimation," *Journal of Software: Evolution and Process,* vol. 31, n° 2, 2018.
S16	M. Hosni, A. Abran and A. Idri, "Improved Effort Estimation of Heterogeneous Ensembles using Filter Selection," em *13th International Conference on Software Technologies*, Porto, Portugal, 2018.
S17	P. Sharma and J. Singh, "Machine Learning Based Effort Estimation Using Standardization," em *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, Greater Noida, India, 2018.
S18	T. R. Benala and R. Mall, "SEET: Software Development Effort Estimation Using Ensemble Techniques," *ACM SIGSOFT Software Engineering Notes,* vol. 43, n° 3, pp. 17–17, 2018.
S19	A. Zakrani and H. Mustapha, "Software Development Effort Estimation Using Random Forests: An Empirical Study and Evaluation," *International Journal of Intelligent Engineering and Systems,* Vols. %1 de %2300 - 311, n° 11, 2018.
S20	O. a. C. K. Malgonde, "An ensemble-based model for predicting agile software development effort," *Empirical Software Engineering,* vol. 24, 04 2019.
S21	S. a. K. S. a. B. P. R. Shukla, "Analyzing Effect of Ensemble Models on Multi-Layer Perceptron Network for Software Effort Estimation," em *2019 IEEE World Congress on Services (SERVICES)*, 2019.
S22	Z. Abdelali, H. Mustapha and N. Abdelwahed, "Investigating the use of random forest in software effort estimation," *Procedia Computer Science,* vol. 148, pp. 343-352, 2019.
S23	P. Phannachitta and K. Matsumoto, "Model-Based Software Effort Estimation–A Robust Comparison of 14 Algorithms widely used in the Data Science Community," *International Journal of Innovative Computing, Information and Control,* vol. 2, 04 2019.
S24	L. Song, L. L. Minku and X. Yao, "Software Effort Interval Prediction via Bayesian Inference and Synthetic Bootstrap Resampling," *Association for Computing Machinery,* vol. 28, p. 46, jan 2019.
S25	W. Amaral, L. Rivero, G. B. Junior and D. Viana, "Using Machine Learning Technique for Effort Estimation in Software Development," *Association for Computing Machinery,* p. 240–245, 2019.
S26	K. Rao and G. Rao, "Ensemble learning with recursive feature elimination integrated software effort estimation: a novel approach," *Evolutionary Intelligence,* vol. 14, 2020.
S27	H. D. P. Carvalho, M. N. C. A. Lima, W. B. Santos and R. A. de A. Fagunde, "Ensemble Regression Models for Software Development Effort Estimation: A Comparative Study," *International Journal of Software Engineering & Applications,* vol. 11, n° 3, p. 71–86, May 2020.
S28	M. A. Shah, D. N. A. Jawawi, M. A. Isa, M. Younas, A. Abdelmaboud and F. Sholichin, "Ensembling Artificial Bee Colony With Analogy-Based Estimation to Improve Software Development Effort Prediction," *IEEE Access,* vol. 8, pp. 58 402-58 415, 2020.
S29	S. Palaniswamy and R. Venkatesan, "Hyperparameters tuning of ensemble model for software effort estimation," *Journal of Ambient Intelligence and Humanized Computing,* pp. 1–11, 2020.
S30	M. Gultekin and O. Kalipsiz, "Story Point-Based Effort Estimation Model with Machine Learning Techniques," *International Journal of Software Engineering and Knowledge Engineering,* vol. 30, n° 01, pp. 43–66, 2020.

Appendix A

See Tables 18 and 19.

Appendix B

See Table 20.

References

Abran, A., 2015. Data Collection and Industry Standards: The ISBSG Repository, Software Project Estimation. John Wiley and Sons, Inc, pp. 161–184.

Cabral, J.T.H.d.A., Oliveira, L.I.A., 2021. Ensemble effort estimation using dynamic selection. J. Syst. Softw. 175, 110904.

Cabral, J.T., Oliveira, A., Nóbrega, J., Araújo, R., 2017. Heterogeneous ensemble dynamic selection for software development effort estimation. In: IEEE 29th International Conference on Tools with Artificial Intelligence. ICTAI, Boston, MA.

Charette, R.N., 2005. Why software fails. IEEE Spectr. 42–49.

Cruz, Rafael M.O., Hafemann, Luiz G., Sabourin, Robert, Cavalcanti, George D., 2020. DESlib: A dynamic ensemble selection library in Python. J. Mach. Learn. Res. 21.

Cruz, Rafael M.O., Sabourin, Robert, Cavalcanti, George D.C., 2018. Dynamic Classifier Selection, Vol. 41. Elsevier Science Publishers B. V., pp. 195–216, no. C.

Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems. London, UK, UK.

Garner, Paul, et al., 2016. When and how to update systematic reviews: consensus and checklist. BMJ 354, i3507. http://dx.doi.org/10.1136/bmj.i3507.

Gultekin, M., Kalipsiz, O., 2020. Story point-based effort estimation model with machine learning techniques. Int. J. Softw. Eng. Knowl. Eng. 30 (01), 43–66.

Hosni, M., Abran, A., Idri, A., 2017. Investigating heterogeneous ensembles with filter feature selection for software effort estimation. In: 12th International Conference on Software Process and Product Measurement. Gothenburg, Sweden.

Hosni, M., Idri, A., 2017. Software effort estimation using classical analogy ensembles based on random subspace. In: Proceedings of the Symposium on Applied Computing. Marrakech, Morocco.

Hou, W.-h., Wang, X.-k., Zhang, H.-y., Wang, J.-q., Li, L., 2020. A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. Knowl.-Based Syst. 208.

Idri, A., Hosni, M., Abran, A., 2016. Systematic literature review of ensemble effort estimation. J. Syst. Softw. 118, 151–175.

Kitchenham, B., Charters, S., 2006. Guidelines for performing systematic literature reviews in software engineering. In: Proceedings of the 28th International Conference on Software Engineering. Shangai, China.

Ko, A.H.R., Sabourin, R., Britto, J.A.S., 2008. From dynamic classifier selection to dynamic ensemble selection. Pattern Recognit. 41 (5), 1718–1731.

Mendes, Emilia, Wohlin, Claes, Felizardo, Katia, Kalinowski, Marcos, 2020. When to update systematic literature reviews in software engineering. J. Syst. Softw. (ISSN: 0164-1212) 167, 110607. http://dx.doi.org/10.1016/j.jss.2020.110607.

Muhammed Niyas, K.P., Thiyagarajan, P., 2021. Alzheimer's classification using dynamic ensemble of classifiers selection algorithms: A performance analysis. Biomed. Signal Process. Control 68.

Shepperd, M., MacDonell, S., 2012. Evaluating prediction systems in software project estimation. Inf. Softw. Technol. 820–827.

Shirabad, J.S., Menzies, T.J., 2005. The PROMISE Repository of Software Engineering Databases. School of Information Technology and Engineering, University of Ottawa, Canada, [Online]. Available: http://promise.site.uottawa.ca/SERepository. [Acesso em 05 jan 2021].

Trendowicz, Adam, Jeffery, Ross, 2014. Software Project Effort Estimation: Foundations and Best Practice Guidelines for Success. Springer International Publishing, Switzerland, http://dx.doi.org/10.1007/978-3-319-03629-8.

Twala, B., Verner, J., 2016. Toward accurate software effort prediction using multiple classifier systems. Comput. Intell. Quant. Softw. Eng. 617, 135–151.

Wen, J., Li, S., Lin, Z., Huc, Y., Huang, C., 2012. Systematic literature review of machine learning based software development effort estimation models. Inf. Softw. Technol. 54 (1), 41–59.

Wohlin, Claes, 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of EASE '14, May 13-14 2014. London, England, http://dx.doi.org/10.1145/2601248.2601268.

Jose Thiago H. de A. Cabral received his B.S. in Computer Science from Federal University of Paraiba in 2007 and the M.Sc. degree in Computer Science from Federal University of Pernambuco, Recife, Brazil in 2012. Currently, he is a Ph.D. candidate in Computer Science at Federal University of Pernambuco. His research interests include Software Effort Estimation, Data Mining and Computacional Intelligence.

Adriano L.I. Oliveira obtained his B.Sc. degree in Electrical Engineering and M.Sc. and Ph.D. degrees in Computer Science from the Federal University of Pernambuco, Brazil, in 1993, 1997 and 2004, respectively. In 2011 he joined the Center for Informatics at Federal University of Pernambuco as an Associate Professor. He was a Visiting Professor at École de Technologie Supérieure (ÉTS, Université du Québec, Montréal, Canadá) from 2018 to 2019. He was an Assistant Professor at Federal Rural University of Pernambuco from 2009 to 2011 and at the Department of Computing Systems of Pernambuco State University from 2002 to 2009. He has published over 140 articles in scientific journals and conferences and one book. He is a Senior Member of the IEEE. His current research interests include deep neural networks, machine learning, pattern recognition, data mining, and applications of these techniques to time series analysis and forecasting, information systems, software engineering, and biomedicine.

Fabio Q.B. da Silva is a professor of software engineering with the Centro de Informática, Universidade Federal de Pernambuco. His research interests include human aspects in software engineering, innovation management, and future thinking. Silva is the leading researcher of the Human Aspects in Software Engineering Research Group (HASE Research) where he conducts field studies with software organizations in Brazil and several other countries with the goal of understanding and improving software practice. He is on the editorial board of Information and Software Technology (IST) and has been the Co-chair of ICSE's Conducting Empirical Studies in Industry Workshop in 2017 and 2018. He was the General Chair of the IEEE/ACM ESEM Symposium in 2019.