

EDCA – An Evolutionary Data-Centric AutoML Framework for Efficient Pipelines

Joana Simões^[0000–0002–6846–3517] and João Correia^[0000–0001–5562–1996]

University of Coimbra, CISUC/LASI – Centre for Informatics and Systems of the
University of Coimbra, Department of Informatics Engineering
`{joanasimoes,jncor}@dei.uc.pt`

Abstract. Automated Machine Learning (AutoML) gained popularity due to the increased demand for Machine Learning (ML) specialists, allowing them to apply ML techniques effortlessly and quickly. AutoML implementations use optimisation methods to identify the most effective ML solution for a given dataset, aiming to improve one or more predefined metrics. However, most implementations focus on model selection and hyperparameter tuning. Despite being an important factor in obtaining high-performance ML systems, data quality is usually an overlooked part of AutoML and continues to be a manual and time-consuming task. This work presents EDCA, an Evolutionary Data Centric AutoML framework. In addition to the traditional tasks such as selecting the best models and hyperparameters, EDCA enhances the given data by optimising data processing tasks such as data reduction and cleaning according to the problems’ needs. All these steps create an ML pipeline that is optimised by an evolutionary algorithm. To assess its effectiveness, EDCA was compared to FLAML and TPOT, two frameworks at the top of the AutoML benchmarks. The frameworks were evaluated in the same conditions using datasets from AMLB classification benchmarks. EDCA achieved statistically similar results in performance to FLAML and TPOT but used significantly less data to train the final solutions. Moreover, EDCA experimental results reveal that a good performance can be achieved using less data and efficient ML algorithm aspects that align with Green AutoML guidelines.

Keywords: Automated Machine Learning · Data-Centric Artificial Intelligence · Green Automated Machine Learning

1 Introduction

Machine Learning (ML) has grown exponentially in the last few years and has been used in various industries. The increased data availability pressures ML experts to create systems to handle it in response to their needs. However, building these ML systems is time-consuming and includes a lot of repetitive tasks. Hence, Automated Machine Learning (AutoML) appears in response to this high demand for experts, promising to reduce the time wasted on creating optimised ML systems [38]. AutoML frameworks use various optimisation algorithms to

identify the most effective model or pipeline for the given ML task [43]. Their main focus has been in the modelling and hyper-parameter tuning phases [25], neglecting other essential pipeline steps [3]. Despite the benefit of automating the search for the best configuration, AutoML has the downside of demanding a high computational power to evaluate all intermediate solutions [36]. Thus, Green Automated Machine Learning (Green AutoML) appeared in response to the current concerns about the computational costs of AutoML. It focuses on reducing the computational power required during the search and, at the same time, in ways to create more energy-efficient solutions.

Additionally, most AutoML frameworks ignore the importance of removing data inconsistencies from the train data and do not optimise the step of the pipeline. Despite data processing is a time-consuming and manual task required before the AutoML optimisation [45,48]. Normally, it is done manually or hard-coded on the frameworks where it is not customised to the data's requirements. Data Centric Artificial Intelligence (DCAI) proposes a shift from the AutoML's model-centric perspective into a data-centric perspective, based on the presumption that improving the data quality increases the accuracy of ML systems [47,4]. Another important data-centric task is to reduce data. Data selection helps to reduce data to only the most relevant one, without losing its information. At the same time, it also reduces the computational costs of training ML systems.

This paper proposes Evolutionary Data Centric AutoML (EDCA), a framework that combines the principles of AutoML, DCAI, and Green AutoML to create a low-cost AutoML framework capable of building efficient ML pipelines for classification problems. It replicates the process made by ML experts to create ML pipelines. It starts by analysing the received data to define the data processing required to handle it. The processing steps required will be integrated into the pipeline structure. Additionally, the pipeline includes steps to select the most relevant data for the problem, both at the instances and features level. This data selection also helps reduce the computational costs during the optimisation. The pipeline ends with a classification model. With the created pipeline structure, an Evolutionary Algorithm (EA) is applied to find its best configuration. Experiments with a subset of the AMLB benchmark datasets [17] indicate that EDCA can achieve simple and efficient solutions for the same predictive performance that use significantly less data.

In this paper, we explored the importance of integrating DCAI techniques in a AutoML framework. In addition, we studied ways to lower the computational costs involved with training based on Green AutoML guidelines. The contributions are the following: (i) proposal of an AutoML framework, EDCA, for the optimisation of a ML pipeline using DCAI techniques to increase data quality and Green AutoML guidelines to reduce computational costs; (ii) experiments with the EDCA and comparison with state-of-the-art AutoML frameworks on benchmark datasets; (iii) analysis of the AutoML frameworks regarding their predictive power, costs (amount of data use), number of solutions evaluated by each one and final solutions achieved; (iv) based on the analysis of experimen-

tal results, EDCA promotes efficient solutions without losing predictive performance, enforcing the ideas of DCAI e Green AutoML.

The remainder of the paper is organised as follows. Section 2 describes the concepts of AutoML, DCAI, and Green AutoML. Then, Section 3 presents EDCA in detail. The experimental study and the results achieved are presented in Section 4. Lastly, Section 5 describes the main conclusions and future work.

2 Related Work

AutoML appeared in response to the high demand of experts to perform ML tasks [18]. Its goal is to reduce the time wasted on repetitive tasks to find the best pipeline for a given ML task [34]. Several implementations are currently working to address AutoML’s challenges. They differ on the parts of the ML pipelines they optimised and on the algorithms used [3]. However, the search spaces increase with additional phases being optimised, making the problem harder [27]. Most frameworks specialise in the model selection and Hyperparameter Optimisation (HPO) phases, with users handling the remaining steps manually, despite occasionally incorporating other phases into their process [25,48].

Arguably one of the first attempts of an AutoML framework, Auto-Weka [34] uses Bayesian optimisation to identify the optimal pipeline and applies meta-learning from past experiments to new datasets. Similarly, Auto-sklearn [14], built on top of Scikit-learn, by including some data-processing steps on the search space to transform the input data and by creating an ensemble with the tested pipelines. However, in later versions, most of the data preprocessing was excluded [13,17]. AutoGluon uses a three-layer stacking ensemble with predefined models, not focusing on hyperparameter tuning [11]. It assumes some models are better, and only trains theoretically best ones when time is limited. AutoGluon also conducts an initial dataset analysis for data processing, but these steps are fixed and not optimised. TPOT, another known framework, represents the ML pipeline in a graph structure and uses genetic programming (GP) [28]. TPOT allows dynamic pipelines created by data preprocessing and predictive operators. However, despite using mechanisms to control pipelines’ complexity, the dynamic pipelines increase the search space because the optimisation must simultaneously identify the optimal structure and configuration of the operators. FEDOT also applies GP to optimise dynamic pipelines that could have more than one predictive model, using a multi-objective evaluation function to control complexity and increase performance. Likewise, AutoML-DSGE [1] optimises dynamic ML pipelines through grammatical evolution and DeepLine [19] applies reinforcement learning to optimise similar dynamic pipelines. Additionally, Wang et al. created FLAML [42] to reduce the computational costs of evaluating several intermediate solutions. The framework focuses on selecting the ideal models, using the estimated cost (CPU time wasted) and performance in previous evaluations to make the decisions. It also uses an incremental instance reduction to lower the costs of the evaluations. More AutoML exists and can be consulted in surveys such as [48,18,2].

AutoML is often criticised for its expensive and resource-consuming search. The frameworks often evaluate similar intermediate solutions without significant improvements in performance. Green AutoML introduces guidelines to create environmentally friendly AutoML frameworks [36]. It identifies effective strategies to minimise resources and lower computational costs usually associated with AutoML. One important guideline is to use less data to test intermediate solutions since their energy consumption and training time are positively correlated with the size of the training data used [41]. Therefore, reducing data is environmentally beneficial and accelerates the optimisation. Additional strategies refer to the use of multi-objective evaluation to incorporate costs on the decision process, and in ways to prevent and overcome performance stagnation [7,36].

Data quality is essential [47] and it influences the quality of the ML systems trained with it [24,23]. Thus, DCAI proposes shifting from the old model-centric approach to handle ML systems into a more data-centric method, where the focus is improving data quality and, respectively, the ML systems. Depending on the data’s needs, various data processing tasks need to be applied such as imputing missing values, removing duplicates, encoding categorical features and normalising data [45]. Despite being time-consuming, few frameworks tried to automate these tasks due to the increase in the search space [45,23,24]. In most AutoML, they are not optimised to the problem’s needs, they are hard-coded or required to be manually handled [48]. Despite recognising data importance, they minimise the problem by creating robust models that handle noisy data [44,27]. Additionally, large datasets can sometimes have lower predictive performance due to increased noise and complexity [26,35,20]. Data Reduction (DR) mitigates this by selecting only relevant instances or features, preserving the original data properties while reducing the computational costs of training ML systems [41,36]. DR techniques include Instance Selection (IS), which narrows data to representative instances (samples), and Feature Selection (FS), which removes redundant features, improving the model’s generalisation ability. EA have been used for both IS [30,6,31,16,37,12] and FS [46,21,33,32], helping to select optimal data subsets. Some works also combine IS and FS [9,10] with EA, but maintain separate populations that are only integrated at the evaluation stage. Additionally, some AutoML frameworks integrate FS into their pipeline, optimising FS algorithms similarly to predictive models [28,34,14]. When IS appears in AutoML frameworks, it is usually a random IS to accelerate the optimisation [42].

To the best of our knowledge, most AutoML frameworks ignore the importance of data and continue with a model-centric perspective. Nevertheless, optimising the data used by the AutoML frameworks has benefits in performance and costs. DCAI showed how removing inconsistencies and transforming data into a more digestible format may increase performance. The necessary preprocessing tasks may vary depending on the problem and should be tailored to each. At the same time, DR techniques aim to select only the best data to help increase performance and the generalisation ability. These DR techniques are also recommended by the Green AutoML community, as they also allow to train more solutions and achieve final solutions with less computational costs associated.

3 EDCA – Evolutionary Data Centric AutoML

EDCA¹ is an AutoML framework capable of creating a complete ML pipeline, attending to the principles of DCAI [47] and Green AutoML [36,41,7], resorting to an automated analysis module for the data combined with an EA that evolves the complete pipeline, model and data. Figure 1 presents EDCA’s process. The input data is first divided into train and validation (1). Then, the ML pipeline steps are created based on an initial train data analysis (3), which determines the preprocessing required. Additionally, the pipeline includes DR steps, which helps to select only the relevant data and reduce the optimisation costs (3). The pipeline structure ends with the optimisation process of a model (3). A Genetic Algorithm (GA) is used to find the best pipeline configuration for the problem within a set time limit (4).

EDCA’s process starts by dividing the input data into train and validation data (Phase 1). The train data passes to Phase 2, whereas the validation data is only used in Phase 4 to assess individuals’ predictive power.

In Phase 2, the data is analysed to assess which processing techniques are required for the given data. It determines the features, types, and characteristics of the data. Each feature in the data could belong to one of four groups: binary, categorical, numerical or identifier. The data characteristics help to understand the types of preprocessing needed for the problem, and each group of features will have a different treatment. After the analysis, a pipeline structure is created (Phase 3). In total, five different preprocessing steps could exist in the pipeline. If the analysis indicates that the data contains numerical features, a normalisation step will be added. The columns with identifiers will be removed since they do not add relevant information. When categorical features exist, an encoder step is added, as the models only digest numerical data. When there are features with missing data, several imputation steps are added to the pipeline, one for each feature type, since the hyperparameters will be different.

Creating the pipeline’s structure based on the data characteristics allows for using only the minimum required preprocessing steps (Phase 3). If all the preprocessing steps were available every time, time would be spent tuning irrelevant methods for the data in question. Thus, using an initial dataset analysis to define the preprocessing required and corresponding search space helps reduce the problem’s complexity and only spend time tuning the essential ML steps. During the optimisation, these processing steps are always present in the individuals.

The ML pipeline includes steps for DR (Phase 3), which allows to select of only the most relevant data for the problem, while, at the same time, reducing the costs and complexity of the problem by accelerating the optimisation. EDCA can include two DR steps for the IS and FS. However, since using them may or may not be beneficial, EDCA uses automatic data optimisation. The automatic data optimisation searches for the best combination of the DR techniques by turning on or off these genes along the optimisation. It can be chosen to use only one type of DR, both, or none of them. Phase 3 also adds a final step to the

¹ EDCA’s GitHub repository

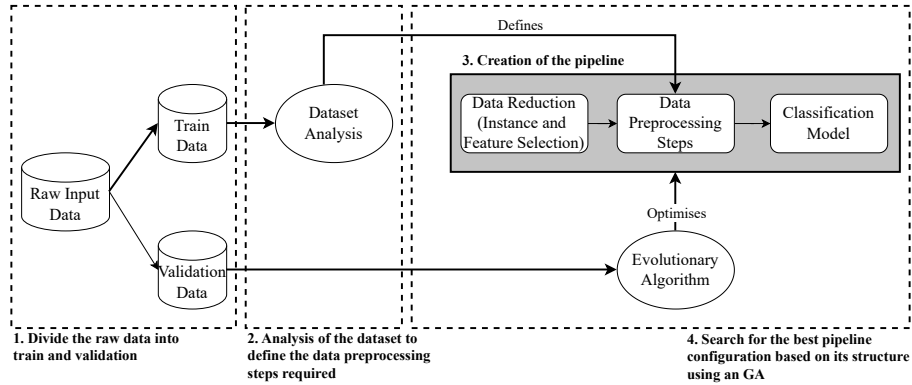


Fig. 1: Process used by EDCA.

pipeline for a predictive model, similar to other frameworks. A suitable model should be used at this stage since it would impact the final results. Overall, the pipeline uses a sequential linear structure, with only one path for data to follow.

To optimise the pipeline, EDCA uses a GA due to its effectiveness in large search spaces [39]. Each individual is represented by a list of genes and each gene represents a step of the ML pipeline, including DR, the different processing required, and a predictive model (see the example of Parent 1 in Figure 2 where it represents the maximum number of genes an individual could have). The DR-related genes are composed of a set of integer numbers corresponding to indices, since in previous works that used a binary representation for the DR the reduction tented to reach only 50% [30]. These DR genes may or not be present on the pipeline since we use a procedure called automatic data optimisation, that can turn on or off their DR genes. When a DR gene is present, it says which indices, related to an instance or feature, should be used to train the pipelines. The indices vary between zero and the maximum length possible, i.e., $[0, max_{instances}]$ or $[0, max_{features}]$. The remaining genes correspond to the other pipeline steps and are composed of the model and its hyperparameters, where each model may vary on the number of hyperparameters. The individuals use a dynamic structure since they may vary depending on the presence or absence of the DR genes, while the others are always present. EDCA implements both crossover and mutation as variation operators.

The crossover operator is applied to interchange genes between individuals. It receives two parents and returns two offspring. The genes of the received individuals are switched between them using uniform crossover, except for the DR-related genes (see Figure 2). The DR genes include a special crossover operator since they use a different representation and they might or may not be present in the individuals. The crossover is applied only if both individuals have the gene. In that case, a one-point crossover occurs at the gene level. If not, the gene is directly inherited from the original individual to the corresponding offspring. The one-point crossover preserves unique values to prevent data augmentation.

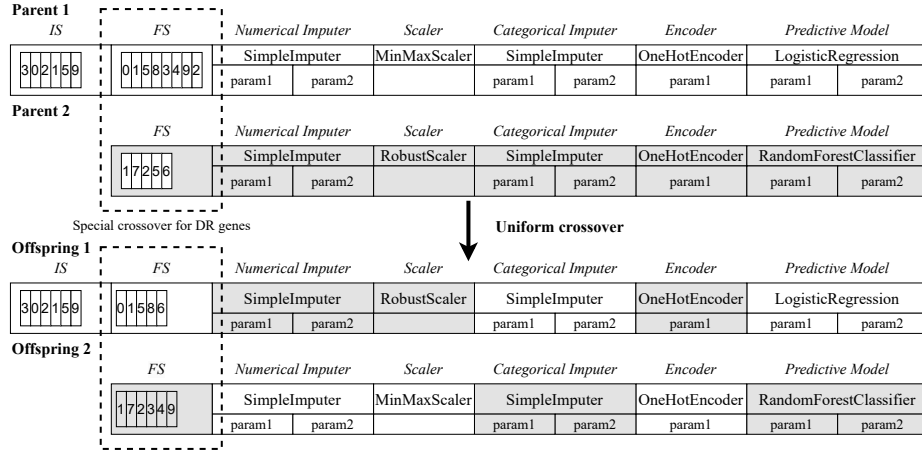


Fig. 2: Example of crossover. It is not applied to the IS genes because they are absent in both individuals. For the FS genes, a gene-level one-point crossover is applied. The non-DR genes use a uniform crossover at the parent level.

Additionally, the mutation operator applied only changes one gene per individual. The general case works by changing the method in a step of the pipeline and its hyperparameters or changing only the hyperparameters, each option with the same likelihood. Since automatic data optimisation is applied to the DR genes, depending on the presence or absence of the gene, different actions may be used (see Figure 3). If the gene is present, it can be deleted or changed. Otherwise, the only option is to add it. When it is chosen to mutate the gene, three different operations can be applied: (i) change $\theta\%$ of the indices for other values; (ii) add $\theta\%$ of new indices; (iii) delete $\theta\%$ of the indices. The user defines the maximum percentage of change from the initial data size, and $\theta\%$ varies between zero and the maximum defined in the runtime.

Each individual's fitness is evaluated using the performance metric error, calculated with the validation data from Phase 1 validation in Figure 1, similar to other frameworks. The fitness ranges from 0 to one ($[0, 1]$), where zero indicates the best performance and one the lowest.

Furthermore, the optimisation uses some techniques to save time and help the search process. To maintain diversity and avoid performance stagnation as recommended by the Green AutoML guidelines [36], the optimisation uses a patience mechanism to wait for improvements on the best solutions, where the population is restarted and new individuals substitute the current population when the performance of the best individuals does not improve after a certain number of generations. The GA also uses elitism to save the best solutions across generations, even when a restart occurs. Similar to other AutoML frameworks, EDCA offers parallel searching to evaluate multiple solutions simultaneously, which allows to accelerate the process and test more generations.

In short, EDCA is composed of different parts (see Figure 1). The entire pipeline consists of an initial DR phase, followed by the preprocessing steps

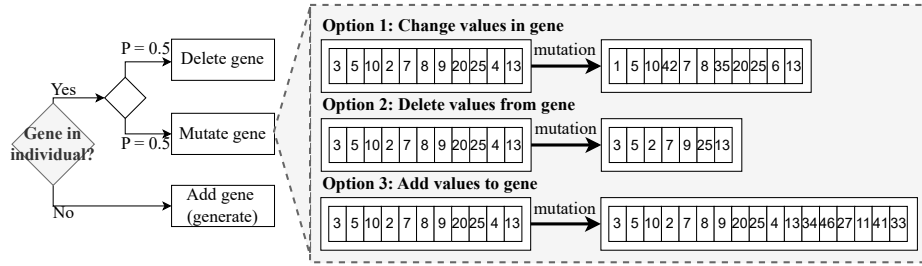


Fig. 3: Flow chart of the mutation applied to DR genes (IS or FS).

dependent on the analysis. Then, the pipeline ends with a predictive model. The pipeline structure is optimised by an GA that explores different configurations during a specified time budget and attempts to optimise a performance metric.

4 Experimental Study

An experimental study was carried out to compare EDCA with state-of-the-art AutoML frameworks, FLAML [42] and TPOT [28]. This section details the setup used (Sub-section 4.1) and the results achieved (Sub-section 4.2).

4.1 Experimental Setup

All the frameworks were evaluated under the same conditions, i.e., they had the same time budget and used the same optimisation metric for the search. The Matthews Correlation Coefficient (MCC), due to being robust to the class used as positive [8,22], was selected to measure the performance during the optimisation and to assess frameworks' predictive ability. Table 1a and 1b present the parameters used for the implementations. EDCA's parameters were defined after evaluating several combinations using train-test divisions in different datasets. Any parameters not listed in Table 1a are set to default for FLAML and TPOT.

EDCA is currently focused on classification problems (binary or multi-class) with tabular data. Therefore, different classification models were considered at this step, with all having different hyperparameters to be tuned. The three frameworks used the same predictive models list, with small variations in the range of values of the hyperparameters. The implementations are from Scikit-Learn [29], LightGBM and XGBoost.

We used a subset of five classification datasets from Gijssbers et al. benchmark study of AutoML frameworks [17]. This subset was selected because the datasets required some data preprocessing work, and for that, each preprocessing type can be applied at least once (see Table 2 with their characteristics).

4.2 Results

For each dataset, 30 independent runs were performed with different random seeds. A 5-fold cross-validation (CV) was applied to each run and the predictive

Table 1: Parameterisation used by the frameworks. FLAML does not use FS, and TPOT does not use IS since they do not implement them.

(a) Common parameters		(b) Evolutionary parameters	
Parameter	Value	Parameter	Value
Runs	30	Probability of Mutation	0.3
External Data Division	CV	Probability of Crossover	0.7
K-fold	5	Elitism Size	1
Stop criterion	Time Budget	Population Size	50
Time Budget (seconds)	900	Tournament Size	3
Optimisation Metric	MCC [8,22]	Patience (no. generations)	5
Instance Selection	True	Automatic Data Optimisation	True
Feature Selection	True	Percentage of change	10%
Parallel Jobs	5		
Internal Data Division	Train-Val		
Validation Percentage	25%		

Table 2: Summary of the datasets. Acronyms: ID: OpenML[40] ID; MV: Missing values; NF: Numerical features; CF: Categorical features; BF: Binary features.

Dataset	ID	Instances	Features	Classes	MV	NF	CF	BF
Australian	40981	690	14	2	No	14	0	0
adult	1590	48842	14	2	Yes	6	7	1
cnae-9	1468	1080	856	9	No	856	0	0
credit-g	31	1000	20	2	No	7	11	0
mfeat-factors	12	2000	216	10	No	216	0	0

performance in each fold was calculated. To verify the statistical significance of the results obtained by EDCA in comparison to the other frameworks, the Mann-Whitney statistical test was employed. A significance level of 99% was used in the statistical tests, using a Bonferroni correction to reduce Type I errors.

The MCC values achieved by the frameworks are presented in Table 3. There are almost no statistically significant differences in MCC values between FLAML and TPOT when compared to EDCA. Additionally, TPOT fails in some datasets when used off-the-shelf without manual data processing.

Considering the percentage of data used in the end, EDCA uses significantly less data than the remaining frameworks regarding total data, instances, and features in almost all the experiments (see Tables 4a, 4b and 4c). This occurs since FLAML only uses incremental sampling to accelerate the optimisation, ignoring FS. With this incremental sampling, it always reaches the maximum data available, i.e., the internal training data available, since 25% is for validation and does not retrain with it. On the other hand, TPOT does not use IS and only applies FS. Considering that EDCA can achieve similar results to FLAML and TPOT with less data, in the context of Green AutoML, it is beneficial to use EDCA due to the lower associated costs for identical predictive performance.

Table 3: Average MCC values (\pm standard deviation) over 30 runs for EDCA, FLAML, and TPOT. Results for EDCA and FLAML after retraining on all data are included, while TPOT always uses all data. Statistically significant differences compared to EDCA are in bold. Arrows show changes after retraining.

Dataset	EDCA	FLAML	TPOT	EDCA-All	FLAML-All
Australian	0.72 \pm 0.01	0.72 \pm 0.02	0.72 \pm 0.02	0.73\pm0.02 ↑	0.72 \pm 0.02
adult	0.63 \pm 0.0	0.63 \pm 0.01	-	0.63\pm0.0	0.63 \pm 0.01
cnae-9	0.93 \pm 0.01	0.93 \pm 0.01	0.94\pm0.01	0.94\pm0.01 ↑	0.93 \pm 0.01
credit-g	0.32 \pm 0.04	0.34 \pm 0.03	-	0.36\pm0.03 ↑	0.34 \pm 0.02
mfeat-factors	0.97 \pm 0.0	0.96\pm0.0	0.97 \pm 0.0	0.97\pm0.0	0.96 \pm 0.0

Table 4: Average percentage of data (\pm standard deviation) over 30 runs used by the frameworks. It includes the percentage of total, instances and features that were calculated by dividing the final data size by the original size. Statistically significant differences compared to EDCA are highlighted in bold.

(a) Percentage of data				(b) Percentage of instances			
Dataset	EDCA	FLAML	TPOT	Dataset	EDCA	FLAML	TPOT
Australian	0.35 \pm 0.06	0.75\pm0.0	1.02\pm0.05	Australian	0.43 \pm 0.09	0.75\pm0.0	1.0\pm0.0
adult	0.64 \pm 0.04	0.75\pm0.0	-	adult	0.66 \pm 0.05	0.75\pm0.0	-
cnae-9	0.59 \pm 0.05	0.75\pm0.0	0.98\pm0.06	cnae-9	0.6 \pm 0.06	0.75\pm0.0	1.0\pm0.0
credit-g	0.4 \pm 0.09	0.75\pm0.0	-	credit-g	0.47 \pm 0.09	0.75\pm0.0	-
mfeat-factors	0.47 \pm 0.06	0.75\pm0.0	0.99\pm0.03	mfeat-factors	0.67 \pm 0.05	0.75\pm0.0	1.0\pm0.0

(c) Percentage of features			
Dataset	EDCA	FLAML	TPOT
Australian	0.84 \pm 0.1	1.0\pm0.0	1.02\pm0.05
adult	0.98 \pm 0.03	1.0\pm0.0	-
cnae-9	0.98 \pm 0.03	1.0\pm0.0	0.98 \pm 0.06
credit-g	0.86 \pm 0.11	1.0\pm0.0	-
mfeat-factors	0.72 \pm 0.1	1.0\pm0.0	0.99\pm0.03

Another relevant finding in EDCA was that when comparing the percentage of instances (Table 4b) and features (Table 4c) achieved, it is shown that the percentage of features is always higher than the percentage of instances, meaning that the features may be more relevant for the final results than the number of instances used. Furthermore, Figure 4 shows the histogram of the DR techniques applied on the final pipelines for each dataset, since they may vary because of the use of automatic data optimisation (see Section 3). The results indicate that EDCA accomplishes very different final solutions, which may related to the fact that it uses the best data available without any constraints. An important finding is that only 7.3% of the final solutions do not use DR and that the most used DR technique is IS alone (52.9%), compared with only using FS which represents

Table 5: Average MCC values (\pm standard deviation) for EDCA, FLAML, TPOT, and FLAML + EDCA and TPOT + EDCA (i.e., FLAML or TPOT retrained with EDCA-selected data). Statistically significant differences from EDCA results are in bold and the "*" reveals statistically significant differences from original framework's MCC values. Arrows show changes after retraining.

Dataset	EDCA	FLAML	TPOT	FLAML+EDCA	TPOT+EDCA
Australian	0.72 ± 0.01	0.72 ± 0.02	0.72 ± 0.02	$0.69 \pm 0.04^*$ ↓	$0.64 \pm 0.1^*$ ↓
adult	0.63 ± 0.0	0.63 ± 0.01	-	0.63 ± 0.0	-
cnae-9	0.93 ± 0.01	0.93 ± 0.01	0.94 ± 0.01	$0.91 \pm 0.01^*$ ↓	$0.92 \pm 0.03^*$ ↓
credit-g	0.32 ± 0.04	0.34 ± 0.03	-	$0.24 \pm 0.07^*$ ↓	-
mfeat-factors	0.97 ± 0.0	0.96 ± 0.0	0.97 ± 0.0	$0.95 \pm 0.0^*$ ↓	$0.96 \pm 0.01^*$ ↓

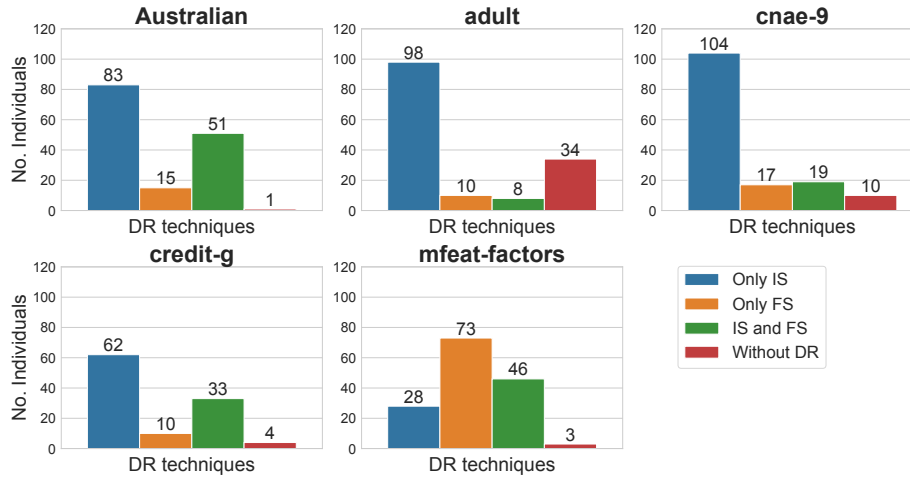


Fig. 4: Distribution of EDCA's DR techniques in a 5-fold CV across 30 runs.

only 17.6% of the cases. This discovery supports the previous point that features are more important so we can reduce more at the instance level.

Additionally, we explored whether using all the data available (train and validation from the CV) would lead to better results, as some studies stated [5,15]. In this case, we only evaluated it for the EDCA and FLAML, as TPOT automatically retrains the pipeline using the entire training dataset at the instance level and does not permit retraining at the feature level. The results only show differences in EDCA when more data is added (see Table 3). When the frameworks are retrained with the entire data, EDCA surpasses the results achieved by the other two frameworks. In addition, the results reveal no differences between the FLAML with all the data and EDCA using only the selected data. Therefore, in the light of Green AutoML, it is better to use EDCA with less data since they achieve identical results. Comparing the FLAML with all training data or only with the selected data, it can be concluded that MCC did not improve when the results were rounded to two decimal points.

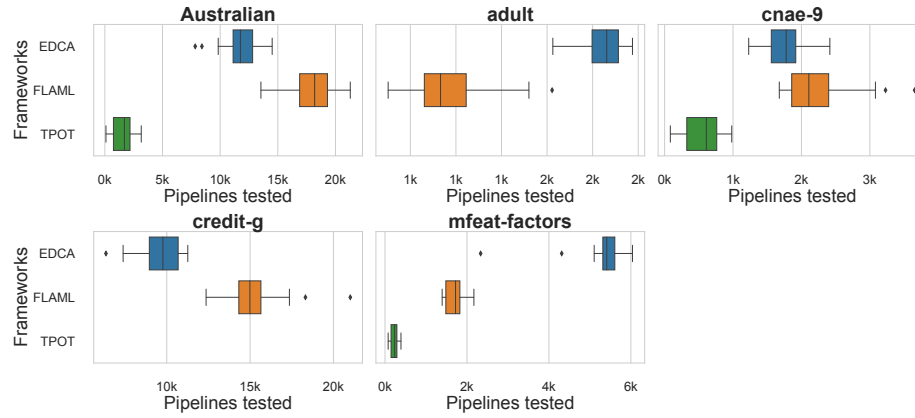


Fig. 5: Number of pipelines evaluated by each framework over 30 runs.

To verify whether combining multiple AutoML frameworks was helpful and whether the data chosen using EDCA was most suitable for the problem, the best pipelines found with FLAML and TPOT retrained with the data optimised by EDCA. Table 5 shows a performance loss when the best solutions on FLAML and TPOT are retrained with EDCA’s data. This loss can be from various sources: (i) the frameworks used different preprocessing optimisations that may influence the results; (ii) the frameworks achieved different final classification models, and the ones achieved by FLAML and TPOT may require more data than the ones achieved with EDCA. In short, optimising the entire pipeline with EDCA achieved good results since the pipeline evolved with all the components, but testing only one part of it may not result in improvements.

Another aspect analysed was how many solutions each framework was able to test. Figure 5 shows that the size of the search space of each framework is inversely related to the number of evaluations made. FLAML has a smaller search space since it only selects the most appropriate classification model and tests more solutions in almost all datasets. Then, EDCA optimises the entire ML pipeline in a linear format, being the second framework that tests more solutions. Finally, TPOT has a larger search space as it represents the solutions as graphs, which are more complex, evaluating, and therefore, fewer solutions. The frameworks that test more solutions, FLAML and EDCA, also apply IS during the optimisation, which accelerates the optimisation.

Additionally, we studied the final solutions each framework created (see Table 6) and concluded that there are differences and the type of solutions can also explain the differences considering the number of evaluations each one had made. EDCA creates complete ML pipelines and, in most cases, it was the only one who had preprocessing steps as predicted, despite TPOT also having those steps on their search space. Despite this, EDCA and TPOT varied more in the final solutions than FLAML. Furthermore, no pattern was found in the distribution of the final models chosen by the frameworks, and they do not agree on the best one (see Figure 6). Also, FLAML tends only to choose gradient-boosting

Table 6: Best pipelines achieved by EDCA, FLAML and TPOT across all runs and folds. "SI" is SimpleImputer. "%I" and "%F" mean the instances and features percentage, respectively. "Imp" is Imputer.

Dataset	Framework	%I	%F	Imp	Scaler	Encoder	Model	MCC
adult	EDCA	0.63	1.0	SI	Standard	One-Hot	XGBClassifier	0.657
adult	FLAML	0.75	1.0	-	-	-	LGBMClassifier	0.658
adult	TPOT	-	-	-	-	-	-	-
Australian	EDCA	0.33	1.0	-	MinMax	-	XGBClassifier	0.853
Australian	FLAML	0.75	1.0	-	-	-	XGBClassifier	0.854
Australian	TPOT	1.0	0.5	-	-	-	LGBMClassifier	0.827
cnae-9	EDCA	0.75	1.0	-	Standard	-	LogisticRegression	0.974
cnae-9	FLAML	0.75	1.0	-	-	-	XGBClassifier	0.969
cnae-9	TPOT	1.0	1.0	-	-	-	StackingEstimator (Extra-TreesClassifier), Standard, LogisticRegression	0.979
credit-g	EDCA	0.5	1.0	-	Robust	One-Hot	LogisticRegression	0.467
credit-g	FLAML	0.75	1.0	-	-	-	LGBMClassifier	0.577
credit-g	TPOT	-	-	-	-	-	-	-
mfeat-factors	EDCA	0.75	0.78	-	Standard	-	LogisticRegression	0.983
mfeat-factors	FLAML	0.75	1.0	-	-	-	XGBClassifier	0.986
mfeat-factors	TPOT	1.0	1.0	-	Standard	-	LogisticRegression	0.989

algorithms. In addition, due to the representation used, TPOT applies ensemble learning in some cases, allowing for more complex solutions. Thus, based on the results in this set of experiments, EDCA generates complete, simpler and efficient solutions trained with less data that, according to the statistical significance tests performed, have similar performance to state-of-the-art frameworks.

5 Conclusion

AutoML has been useful to accelerate the development of ML systems. However, despite knowing that data quality impacts the accuracy of the ML, most AutoML frameworks efforts only focus on predictive model tuning. Additionally, AutoML is criticised for its resource-intense optimisation. Therefore, in this paper, we proposed a data-centric AutoML framework called EDCA - Evolutionary Data Centric AutoML, focused on improving the entire ML classification pipeline from data processing to model tuning, while also reducing the data to lower costs. EDCA starts by analysing the given dataset to infer which data transformations must be applied. Considering the costs associated with training the ML systems, it includes a component of automatic data optimisation to select the most relevant data (instances and features) for the training. Using DR during the optimisation and in the final pipelines allows for the reduction of the associated costs. Like other AutoML frameworks, EDCA also selects the best predictive model for the problem and for this end, a GA optimises the pipeline.

EDCA was compared to state-of-the-art frameworks, FLAML and TPOT, on benchmark datasets. When evaluated on the same conditions, EDCA achieved similar predictive performance results to others. The reduction and transformation of the received data did not put EDCA at a disadvantage against FLAML

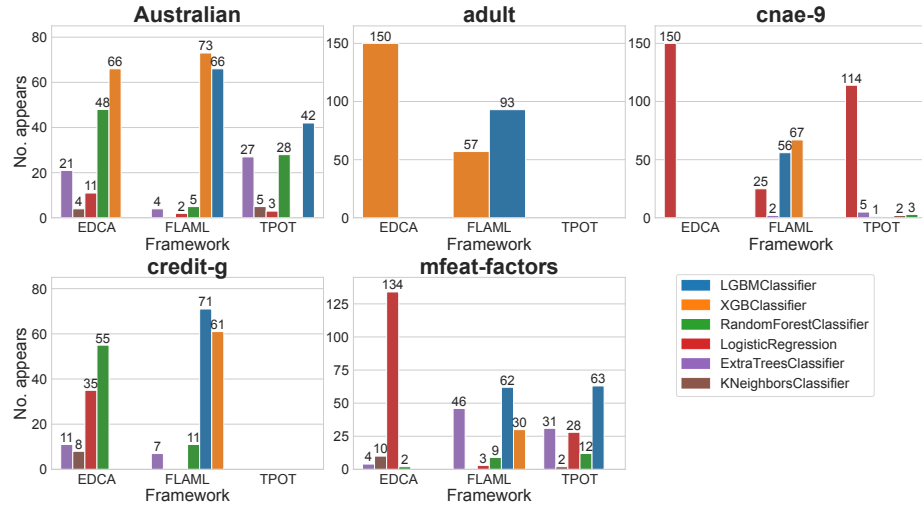


Fig. 6: Distribution of the frameworks' best models in a 5-fold CV over 30 runs.

and TPOT. EDCA uses significantly less data on the final solutions, which in addition to the MCC values achieved by the three frameworks, shows that EDCA is a low-cost AutoML capable of creating simpler but efficient ML solutions. This positive outcome indicates that it is possible to accomplish good results with low costs by optimising the data. In summary, given the concerns around resource-intensive AutoML frameworks, based on the results attained so far suggest that EDCA offers a more efficient and cost-effective AutoML solution, aligning with the principles of Green AutoML.

In the future, we will be conducting experiments about fitness function's impact on the final ML pipelines and the impact of DR on the evolutionary stage in terms of costs, while incorporating more Green AutoML guidelines. Also, we want to make a deeper study about the EA parameters used, which were obtained through empirical testing. Finally, building on our initial results, we aim to extend the comparison with literature by testing EDCA against other AutoML frameworks. Furthermore, EDCA will also be evaluated on more benchmark datasets and real-world datasets to see its generalisation capabilities.

Acknowledgements This work has been partially supported by Project "NEXUS Pacto de Inovação – Transição Verde e Digital para Transportes, Logística e Mobilidade". ref. No. 7113, supported by the Recovery and Resilience Plan (PRR); by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI; by the European Funds Next Generation EU, following Notice No. 02/C05-i01/2022.PC645112083-00000059 (project 53), Component 5 - Capitalization and Business Innovation - Mobilizing Agendas for Business Innovation; by the FCT - Fundação para a Ciência e a Tecnologia, I.P., in the framework of the Project UIDB/00326/2025 and UIDP/00326/2025.

References

1. Assunção, F., Lourenço, N., Ribeiro, B., Machado, P.: Evolution of scikit-learn pipelines with dynamic structured grammatical evolution. In: Castillo, P.A., Laredo, J.L.J., de Vega, F.F. (eds.) *Applications of Evolutionary Computation - 23rd European Conference, EvoApplications 2020, Held as Part of EvoStar 2020, Seville, Spain, April 15-17, 2020, Proceedings. Lecture Notes in Computer Science*, vol. 12104, pp. 530–545. Springer (2020). https://doi.org/10.1007/978-3-030-43722-0_34
2. Baratchi, M., Wang, C., Limmer, S., van Rijn, J.N., Hoos, H.H., Bäck, T., Olhofer, M.: Automated machine learning: past, present and future. *Artif. Intell. Rev.* **57**(5), 122 (2024). <https://doi.org/10.1007/S10462-024-10726-1>
3. Barbudo, R., Ventura, S., Romero, J.R.: Eight years of automl: categorisation, review and trends. *Knowl. Inf. Syst.* **65**(12), 5097–5149 (2023). <https://doi.org/10.1007/S10115-023-01935-1>
4. Bilal, M., Ali, G., Iqbal, M.W., Anwar, M., Malik, M.S.A., Kadir, R.A.: Auto-prep: Efficient and automated data preprocessing pipeline. *IEEE Access* **10**, 107764–107784 (2022). <https://doi.org/10.1109/ACCESS.2022.3198662>
5. Boonyanunta, N., Zeephongsekul, P.: Predicting the relationship between the size of training sample and the predictive power of classifiers. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) *Knowledge-Based Intelligent Information and Engineering Systems, 8th International Conference, KES 2004, Wellington, New Zealand, September 20-25, 2004. Proceedings. Part III. Lecture Notes in Computer Science*, vol. 3215, pp. 529–535. Springer (2004). https://doi.org/10.1007/978-3-540-30134-9_71
6. Cano, J.R., Herrera, F., Lozano, M.: Using evolutionary algorithms as instance selection for data reduction in kdd: an experimental study. *IEEE Trans. Evol. Comput.* **7**(6), 561–575 (2003). <https://doi.org/10.1109/TEVC.2003.819265>
7. Castellanos-Nieves, D., García-Forte, L.: Improving automated machine-learning systems through green ai. *Applied Sciences* **13**(20), 11583 (2023)
8. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* **21**, 1–13 (2020)
9. De Souza, J.T., Do Carmo, R.A.F., De Campos, G.A.L.: A novel approach for integrating feature and instance selection. In: 2008 International Conference on Machine Learning and Cybernetics. vol. 1, pp. 374–379 (2008). <https://doi.org/10.1109/ICMLC.2008.4620434>
10. Derrac, J., García, S., Herrera, F.: A first study on the use of coevolutionary algorithms for instance and feature selection. In: Corchado, E., Wu, X., Oja, E., Álvaro Herrero, Baroque, B. (eds.) *Hybrid Artificial Intelligence Systems, 4th International Conference, HAIS 2009, Salamanca, Spain, June 10-12, 2009. Proceedings. Lecture Notes in Computer Science*, vol. 5572, pp. 557–564. Springer (2009). https://doi.org/10.1007/978-3-642-02319-4_67
11. Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., Smola, A.J.: Autogluon-tabular: Robust and accurate automl for structured data. *CoRR abs/2003.06505* (2020)
12. Fernandes, E.R.Q., de Carvalho, A.C.P.L.F.: Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. *Inf. Sci.* **494**, 141–154 (2019). <https://doi.org/10.1016/J.INS.2019.04.052>

13. Feurer, M., Eggenberger, K., Falkner, S., Lindauer, M., Hutter, F.: Auto-sklearn 2.0: Hands-free automl via meta-learning. *J. Mach. Learn. Res.* **23**, 261:1–261:61 (2022). <https://jmlr.org/papers/v23/21-0992.html>
14. Feurer, M., Klein, A., Eggenberger, K., Springenberg, J.T., Blum, M., Hutter, F.: Auto-sklearn: Efficient and robust automated machine learning. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds.) *Automated Machine Learning - Methods, Systems, Challenges*, pp. 113–134. The Springer Series on Challenges in Machine Learning, Springer (2019). https://doi.org/10.1007/978-3-030-05318-5_6
15. Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., Ngo, L.H.: Predicting sample size required for classification performance. *BMC Medical Informatics Decis. Mak.* **12**, 8 (2012). <https://doi.org/10.1186/1472-6947-12-8>
16. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evol. Comput.* **17**(3), 275–306 (2009). <https://doi.org/10.1162/EVCO.2009.17.3.275>
17. Gijbbers, P., Bueno, M.L.P., Coors, S., LeDell, E., Poirier, S., Thomas, J., Bischl, B., Vanschoren, J.: Amlb: an automl benchmark. *J. Mach. Learn. Res.* **25**, 101:1–101:65 (2024)
18. He, X., Zhao, K., Chu, X.: Automl: A survey of the state-of-the-art. *Knowl. Based Syst.* **212**, 106622 (2021). <https://doi.org/10.1016/J.KNOSYS.2020.106622>
19. Heffetz, Y., Vainshtein, R., Katz, G., Rokach, L.: Deepline: Automl tool for pipelines generation using deep reinforcement learning and hierarchical actions filtering. In: Gupta, R., Liu, Y., Tang, J., Prakash, B.A. (eds.) *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23–27, 2020*. pp. 2103–2113. ACM (2020). <https://doi.org/10.1145/3394486.3403261>
20. Hoens, T.R., Chawla, N.V.: Imbalanced Datasets: From Sampling to Classifiers. In: *Imbalanced Learning*, pp. 43–59. John Wiley & Sons, Ltd (2013). <https://doi.org/10.1002/9781118646106.ch3>
21. Huang, B.Q., Buckley, B., Kechadi, M.T.: Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Expert Syst. Appl.* **37**(5), 3638–3646 (2010). <https://doi.org/10.1016/J.ESWA.2009.10.027>
22. Jurman, G., Riccadonna, S., Furlanello, C.: A comparison of mcc and cen error measures in multi-class prediction. *PLOS ONE* **7**(8), 1–8 (08 2012). <https://doi.org/10.1371/journal.pone.0041882>, <https://doi.org/10.1371/journal.pone.0041882>
23. Krishnan, S., Wang, J., Franklin, M.J., Goldberg, K., Kraska, T., Milo, T., Wu, E.: Sampleclean: Fast and reliable analytics on dirty data. *IEEE Data Eng. Bull.* **38**(3), 59–75 (2015)
24. Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., Zhang, C.: Cleanml: A study for evaluating the impact of data cleaning on ML classification tasks. In: *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19–22, 2021*. pp. 13–24. IEEE (2021). <https://doi.org/10.1109/ICDE51399.2021.00009>
25. Majidi, F., Openja, M., Khomh, F., Li, H.: An empirical study on the usage of automated machine learning tools. In: *IEEE International Conference on Software Maintenance and Evolution, ICSME 2022, Limassol, Cyprus, October 3–7, 2022*. pp. 59–70. IEEE (2022). <https://doi.org/10.1109/ICSME55016.2022.00014>
26. Nalepa, J., Myller, M., Piechaczek, S., Hrynczenko, K., Kawulok, M.: Genetic selection of training sets for (not only) artificial neural networks. In: Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds.) *Beyond Databases, Architectures and Structures. Facing the Challenges of Data*

- Proliferation and Growing Variety - 14th International Conference, BDAS 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 18-20, 2018, Proceedings. Communications in Computer and Information Science, vol. 928, pp. 194–206. Springer (2018). https://doi.org/10.1007/978-3-319-99987-6_15
27. Neutatz, F., Chen, B., Alkhatib, Y., Ye, J., Abedjan, Z.: Data cleaning and automl: Would an optimizer choose to clean? *Datenbank-Spektrum* **22**(2), 121–130 (2022). <https://doi.org/10.1007/S13222-022-00413-2>
 28. Olson, R.S., Bartley, N., Urbanowicz, R.J., Moore, J.H.: Evaluation of a tree-based pipeline optimization tool for automating data science. In: Friedrich, T., Neumann, F., Sutton, A.M. (eds.) *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*, Denver, CO, USA, July 20 - 24, 2016. pp. 485–492. ACM (2016). <https://doi.org/10.1145/2908812.2908918>
 29. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011). <https://doi.org/10.5555/1953048.2078195>
 30. Rathee, S., Ratnoo, S., Ahuja, J.: Instance Selection Using Multi-objective CHC Evolutionary Algorithm. In: Fong, S., Akashe, S., Mahalle, P.N. (eds.) *Information and Communication Technology for Competitive Strategies*. pp. 475–484. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-0586-3_48
 31. Reeves, C.R., Bush, D.R.: Using Genetic Algorithms for Training Data Selection in RBF Networks. In: Liu, H., Motoda, H. (eds.) *Instance Selection and Construction for Data Mining*, pp. 339–356. Springer US, Boston, MA (2001). https://doi.org/10.1007/978-1-4757-3359-4_19
 32. Souza, F., Matias, T., Araújo, R.: Co-evolutionary genetic multilayer perceptron for feature selection and model design. In: Mammeri, Z. (ed.) *IEEE 16th Conference on Emerging Technologies & Factory Automation, ETFA 2011, Toulouse, France, September 5-9, 2011*. pp. 1–7. IEEE (2011). <https://doi.org/10.1109/ETFA.2011.6059084>
 33. Tan, C.J., Lim, C.P., Cheah, Y.: A multi-objective evolutionary algorithm-based ensemble optimizer for feature selection and classification with neural network models. *Neurocomputing* **125**, 217–228 (2014). <https://doi.org/10.1016/J.NEUCOM.2012.12.057>, <https://doi.org/10.1016/j.neucom.2012.12.057>
 34. Thornton, C., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Auto-weka: combined selection and hyperparameter optimization of classification algorithms. In: Dhillon, I.S., Koren, Y., Ghani, R., Senator, T.E., Bradley, P., Parekh, R., He, J., Grossman, R.L., Uthurusamy, R. (eds.) *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. pp. 847–855. ACM (2013). <https://doi.org/10.1145/2487575.2487629>
 35. Tong, L.I., Chang, Y.C., Lin, S.H.: Determining the optimal re-sampling strategy for a classification model with imbalanced data using design of experiments and response surface methodologies. *Expert Syst. Appl.* **38**(4), 4222–4227 (2011). <https://doi.org/10.1016/J.ESWA.2010.09.087>
 36. Tornede, T., Tornede, A., Hanselle, J., Wever, M., Mohr, F., Hüllermeier, E.: Towards green automated machine learning: Status quo and future directions. *CoRR abs/2111.05850* (2021)
 37. Triguero, I., Galar, M., Merino, D., Maillo, J., Bustince, H., Herrera, F.: Evolutionary undersampling for extremely imbalanced big data classification under apache

- spark. In: IEEE Congress on Evolutionary Computation, CEC 2016, Vancouver, BC, Canada, July 24-29, 2016. pp. 640–647. IEEE (2016). <https://doi.org/10.1109/CEC.2016.7743853>
38. Truong, A., Walters, A., Goodsitt, J., Hines, K.E., Bruss, C.B., Farivar, R.: Towards automated machine learning: Evaluation and comparison of automl approaches and tools. In: 31st IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2019, Portland, OR, USA, November 4-6, 2019. pp. 1471–1479. IEEE (2019). <https://doi.org/10.1109/ICTAI.2019.00209>
39. Tsai, C.F., Eberle, W., Chu, C.Y.: Genetic algorithms in feature and instance selection. *Knowl. Based Syst.* **39**, 240–247 (2013). <https://doi.org/10.1016/J.KNOSYS.2012.11.005>
40. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: networked science in machine learning. *CoRR abs/1407.7722* (2014)
41. Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., Hotellier, E.: Data-centric green ai an exploratory empirical study. In: International Conference on ICT for Sustainability, ICT4S 2022, Plovdiv, Bulgaria, June 13-17, 2022. pp. 35–45. IEEE (2022). <https://doi.org/10.1109/ICT4S55073.2022.00015>
42. Wang, C., Wu, Q., Weimer, M., Zhu, E.: Flaml: A fast and lightweight automl library. In: Smola, A., Dimakis, A., Stoica, I. (eds.) *Proceedings of the Fourth Conference on Machine Learning and Systems, MLSys 2021*, virtual, April 5-9, 2021. mlsys.org (2021)
43. Wang, D., Weisz, J.D., Muller, M.J., Ram, P., Geyer, W., Dugan, C., Tausczik, Y.R., Samulowitz, H., Gray, A.G.: Human-ai collaboration in data science: Exploring data scientists’ perceptions of automated ai. *Proc. ACM Hum. Comput. Interact.* **3**(CSCW), 211:1–211:24 (2019). <https://doi.org/10.1145/3359313>
44. Whang, S.E., Roh, Y., Song, H., Lee, J.: Data collection and quality challenges in deep learning: a data-centric AI perspective. *VLDB J.* **32**(4), 791–813 (2023). <https://doi.org/10.1007/S00778-022-00775-9>, <https://doi.org/10.1007/s00778-022-00775-9>
45. Xin, D., Wu, E.Y., Lee, D.J.L., Salehi, N., Parameswaran, A.G.: Whither automl? understanding the role of automation in machine learning workflows. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S.M. (eds.) *CHI ’21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. pp. 83:1–83:16. ACM (2021). <https://doi.org/10.1145/3411764.3445306>
46. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. *IEEE Trans. Evol. Comput.* **20**(4), 606–626 (2016). <https://doi.org/10.1109/TEVC.2015.2504420>
47. Zha, D., Bhat, Z.P., Lai, K., Yang, F., Jiang, Z., Zhong, S., Hu, X.: Data-centric artificial intelligence: A survey. *CoRR abs/2303.10158* (2023). <https://doi.org/10.48550/ARXIV.2303.10158>
48. Zöllner, M.A., Huber, M.F.: Benchmark and survey of automated machine learning frameworks. *J. Artif. Intell. Res.* **70**, 409–472 (2021). <https://doi.org/10.1613/JAIR.1.11854>