Empirical Software Engineering (2021) 26: 53
https://doi.org/10.1007/s10664-020-09906-8

How to Better Distinguish Security Bug Reports (Using Dual Hyperparameter Optimization)

Rui Shu[1] · Tianpei Xia[1] · Jianfeng Chen[1] · Laurie Williams[1] · Tim Menzies[1]

Accepted: 6 November 2020 / Published online: 5 April 2021

Abstract

Background In order that the general public is not vulnerable to hackers, security bug reports need to be handled by small groups of engineers before being widely discussed. But learning how to distinguish the security bug reports from other bug reports is challenging since they may occur rarely. Data mining methods that can find such scarce targets require extensive optimization effort.

Goal The goal of this research is to aid practitioners as they struggle to optimize methods that try to distinguish between rare security bug reports and other bug reports.

Method Our proposed method, called SWIFT, is a *dual optimizer* that optimizes *both* learner and pre-processor options. Since this is a large space of options, SWIFT uses a technique called ϵ*-dominance* that learns how to avoid operations that do not significantly improve performance.

Result When compared to recent state-of-the-art results (from FARSEC which is published in TSE'18), we find that the SWIFT's dual optimization of both pre-processor and learner is more useful than optimizing each of them individually. For example, in a study of security bug reports from the Chromium dataset, the median recalls of FARSEC and SWIFT were 15.7% and 77.4%, respectively. For another example, in experiments with data from the Ambari project, the median recalls improved from 21.5% to 85.7% (FARSEC to SWIFT).

Conclusion Overall, our approach can quickly optimize models that achieve better recalls than the prior state-of-the-art. These increases in recall are associated with moderate increases in false positive rates (from 8% to 24%, median). For future work, these results suggest that dual optimization is both practical and useful.

Keywords Hyperparameter Optimization · Data pre-processing · Security bug report

Communicated by: Bram Adams

✉ Tim Menzies
timm@ieee.org

Extended author information available on the last page of the article.

1 Introduction

Security bug detection is a pressing current concern. A report from NIST comments that "Current systems perform increasingly vital tasks and are widely known to possess vulnerabilities" (Black et al. 2016) (and by "vulnerability", they mean a weakness in the computational logic (e.g., code) found in software and some hardware components (e.g., firmware) that, when exploited, results in a negative impact on confidentiality, integrity, or availability (MITRE 2017)). Daily, news reports reveal increasingly sophisticated security breaches. As seen in those reports, a single vulnerability can have devastating effects. For example, a data breach of Equifax caused the personal information of as many as 143 million Americans – or nearly half the country – to be compromised (2019). The WannaCry ransomware attack (2017) crippled British medical emergency rooms, delaying medical procedures for many patients.

Developers capture and document software bugs and issues into bug reports which are submitted to bug tracking systems. For example, the Mozilla bug database maintains more than 670,000 bug reports with 135 new bug reports added each day (Chen and et al 2013). Submitted bug reports are explicitly labeled as a security bug report (SBR) or non-security bug report (NSBR). Within such bug tracking systems, Peters et al. (2018) warn that it is crucial to correctly identify security bug reports and distinguish them from other non-security bug reports. They note that software vendors ask that security bug reports should be reported directly and privately to their own engineers. These engineers then assess the bug reports and, when necessary, offer a security patch. The security bug, and its associated patch, can then be documented and disclosed via public bug tracking systems. This approach maximizes the probability that a patch is widely available before hackers exploit a vulnerability. However, due to the lack of security expertise knowledge, bug reporters sometimes mislabel security bug reports as non-security bug reports (Gegick et al. 2010). There are cases when they are not sure when their bug is a non-security bug (which can be safely disclosed) or when that bug is a security bug (that needs to be handled more discretely). For example, Fig. 1 demonstrates a security bug report from the Apache Ambari project, which is mislabelled as non-security bug report. It is a labor intensive process and thus impractical for security practitioners to identify mislabelled security bug reports within a large set of thousands of other non-security bug reports.

The problem that researchers need to address is how to distinguish security bug reports properly. To tackle this problem, researchers have adopted various techniques. One technique is to apply text mining to the security bug reports (Gegick et al. 2010; Goseva-Popstojanova and Tyo 2018; Xia et al. 2014, 2016). The main idea here is to find some combinations of relevant keywords in the bug reports (as well as features such as word frequency) which are then combined together into classification models. But learning such models is a challenging task since the ratio of security bug reports to other kinds of bug reports may be very low. For example, data sets from Peters et al. (2018) show among the 45,940 bug reports, only 0.8% are security bug reports. Various methods exist for mining such rarefied data – but those methods require extensive optimization effort before they work well on a particular data set. Peters et al. proposed FARSEC (Peters et al. 2018), a text mining method that used irrelevancy pruning (i.e., filtering). In their approach, developers first identified security related words. Next, they pruned away the irrelevant bug reports (where "irrelevant" means "does not have those security-related keywords"). FARSEC was evaluated using bug reports from one Chromium project and four Apache projects.

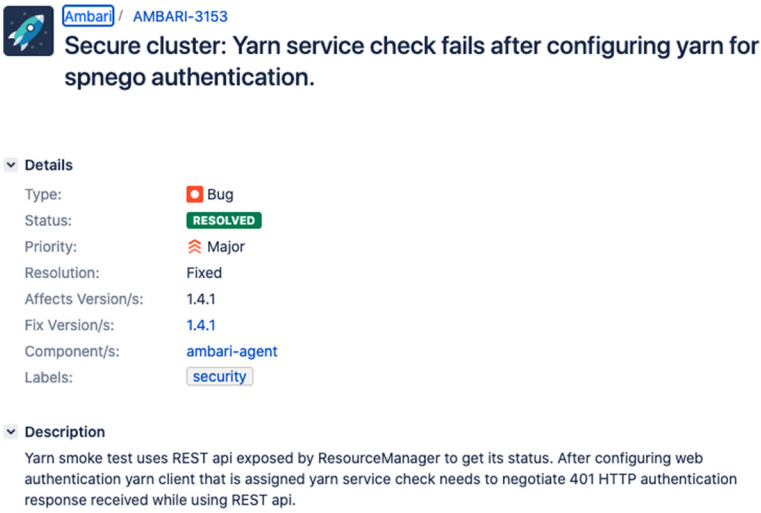

Fig. 1 An example of security bug report from the Apache Ambari project mislabelled as non-security bug report from Peters et al. (2018)

The conjecture of this paper is that this text mining-based method for security bug reports (e.g. as done with FARSEC) can be further enhanced. For example, FARSEC applied its data miners using their default "off-the-shelf" configurations. Recently it has been shown that *hyperparameter optimization* (which automatically learns the "magic" control parameters of an algorithm) can result in better learners that outperform the learners with "off-the-shelf" configurations (Agrawal et al. 2018; Agrawal and Menzies 2018; Fu et al. 2016; Herodotou et al. 2011; Tantithamthavorn et al. 2016; Van Aken et al. 2017; Menzies and Shepperd 2019). To the best of our knowledge, this paper is the first attempt to apply hyperparameter optimization to learn models that better distinguish security bug reports. To that end, we separate and apply three different kinds of optimization strategies:

1. *Learner* hyperparameter optimization to adjust the parameters of the data miner; e.g., how many trees to use in random forest, or what values to use in the kernel of Support Vector Machine (SVM).
2. *Pre-processor* hyperparameter optimization to adjust any adjustment to the training data, prior to learning; e.g., to learn how to control outlier removal or, how to handle the class imbalance problem.
3. *Dual* hyperparameter optimization that combines 1 and 2.

Standard practice in the search-based SE literature explores just learner or pre-processor options, but seldom both. There are good reasons for this – the space of hyperparameters is exponential on the number of optimization options. Hence optimizing *both* the learner *and* pre-processor options is an exponentially slow process. Nevertheless, this paper shows that if dual optimization can terminate, then it is a useful method. For example, for distinguishing security bug reports, dual optimization performs better than just optimizing learner or pre-processor options individually. This paper succeeds at dual optimization, despite its exponential nature, uses a technique called ϵ*-dominance* to ignore operations that do not significantly improve the performance. We call this method SWIFT in our work.

In order to demonstrate the efficiency of dual optimization (i.e., SWIFT), we made comparison experiments with the baseline approach (i.e., FARSEC) as well as state-of-the-art individual optimization methods (i.e., optimizing learners or optimizing pre-processors with the differential evolutionary algorithm). To make that demonstration, we apply dual hyper-parameter optimization to the options of Table 1. We make no claim that this is the entire set of possible options. Rather, we just say that (a) any reader of the recent SE data mining literature might have seen many of these; (b) that reader might be tempted to try optimizing

Table 1 List of pre-processors and learners explored in this study

Type	Name	Description
Pre-processor	Normalizer	Normalize samples individually to unit norm.
	StandardScalar	Standardize features by removing the mean and scaling to unit variance.
	MinMaxScaler	Transforms features by scaling each feature to a given range.
	MaxAbsScaler	Scale each feature by its maximum absolute value.
	RobustScalar	Scale features using statistics that are robust to outliers.
	KernelCenterer	Center a kernel matrix.
	QuantileTransformer	Transform features using quantiles information.
	PowerTransformer	Apply a power transform featurewise to make data more Gaussian-like.
	Binarizer	Binarize data (set feature values to 0 or 1) according to a threshold.
	PolynominalFeatures	Generate polynomial and interaction features.
	SMOTE	Synthetic Minority Over-sampling Technique.
Learner	Random Forest (RF)	Generate conclusions using multiple entropy-based decision trees.
	K Nearest Neighbors (KNN)	Classify a new instance by finding "K" examples of similar instances.
	Naive Bayes (NB)	Classify a new instance by (a) collecting mean and standard deviations of attributes in old instances of different classes; (b) returning the class whose attributes are statistically most similar to the new instance.
	Logistic Regression (LR)	Map the output of a regression into $0 \leq n \leq 1$; thus enabling using regression for classification.
	Multilayer Perceptron (MLP)	A deep artificial neural network which is composed of more than one perceptron.

Standard practice in previous literature is to optimize none or just one of these two groups (Bennin et al. 2019; Agrawal et al. 2018; Agrawal and Menzies 2018; Fu et al. 2016; Tantithamthavorn et al. 2018). Note that a dual optimizer simultaneously explores both learner and pre-processing options

Note: The listed pre-processors and learners are based on scikit-learn version 0.21.2. SMOTE is implemented independently without using existing scikit-learn library.

the Table 1 options; (c) when we optimize these options in our method, we found that our models were better than the prior state-of-the-art (Peters et al. 2018).

This study is structured around the following research questions:

RQ1. Can hyperparameter optimization techniques improve the performance of models that better distinguish security bug reports from other bug reports?

We find that the dual hyperparameter optimization approach better distinguishes security bug reports from non-security bug reports. Specifically, our new method increases the recall on the security bug reports from 21.5% to 66.7% (median values for FARSEC and SWIFT, respectively). This recall increase is associated with moderate false alarm rate increase from 8.0% to 24.0% (median values, FARSEC to SWIFT).

RQ2. When learning how to distinguish security bug reports, is it better to dual optimize the learners and the data pre-processors?

We will show that dual optimization is statistically significantly better in 31/40 data sets with regard to recall results. This is more than twice as many wins as other approaches explored in this paper. In addition, the dual optimization used here is faster (and scales better to more complex problems) than other techniques.

RQ3. Can hyperparameter optimization further improve the performance of ranking security bug reports?

From the ranking evaluation experiment results, we can observe that individual hyperparameter optimization can achieve better ranking score than the best filter treatment from FARSEC for all five projects studied here. In addition, dual optimization is better than individual optimization in this metric across all five projects.

In summary, the contributions of this paper are:

- *An improved result on prior state-of-the-art.* Specifically, to distinguish security bug reports from non-security bug reports, our methods are better than those reported in the previous FARSEC paper from TSE'18.
- *A comment on the value of optimizing (a) data pre-processors or (b) data mining learners.* Specifically, to identify rare events, we show that dual optimization of (a) and (b) does much better than optimizing either, individually.
- *A demonstration of the practicality of dual optimization.* As shown below, the overall runtime for dual optimization (i.e., SWIFT) is five minutes for small datasets and 12 minutes for larger datasets such as the Chromium project on average. This is an important result since our pre-experimental concern is that the cross-product of the option space between the (a) data pre-processors and (b) data mining learners would be so large as to preclude dual optimization.

The remainder of this paper is organized as follows. We introduce research background and related work in Section 2. We then describe the details of our approach in Section 3. In Section 4, we present our experiment details, including hyperparameter optimization ranges, datasets, experiment rig, and metrics, etc. We answer proposed research questions in

Section 5. We deliver the take-away messages in Section 6 and discuss the threats to validity in Section 7 and then conclude in Section 8.

2 Background and Related Work

Various methods have been applied to address the need for more secure software. This section first discusses how data mining has been applied to this problem, then we introduce the state-of-the-art FARSEC technique, after which we introduce more details of hyperparameter optimization.

2.1 Security Bug Reports and Data Mining

Data mining has recently been widely applied in bug report analysis, such as identification of duplicated bug reports (Sun et al. 2011; Lazar et al. 2014; Hindle et al. 2016; Deshmukh et al. 2017), prediction of the severity or impact of a reported bug (Lamkanfi et al. 2010; Zhang et al. 2015; Tian et al. 2012; Yang et al. 2016; Yang et al. 2017), extraction of execution commands and input parameters from performance bug reports (Han et al. 2018), assignment of the priority labels to bug reports (Tian et al. 2015), bug report field reassignment and refinement prediction (Xia et al. 2016) and identify vulnerabilities from commit message and bug reports (Zhou and Sharma 2017) .

In particular, a few studies of bug report classification are more relevant to our work. Some of those approaches focus on building bug classification models based on analyzing bug reports with text mining. For example, Zhou et al. (2016) leveraged text mining techniques, analyzed the summary parts of bug reports and fed into machine learning learners. Xia et al. (2014) developed a framework that applied text mining technology on bug reports and trained a model on bug reports with known labels (i.e., configuration or non-configuration). The trained model was used to predict the new bug reports. Goseva-Popstojanova and Tyo (2018) used different types of textual feature vectors and focused on applying both supervised and unsupervised algorithms in classifying security and non-security related bug reports. Wijayasekara et al. (2014) extracted textual information by utilizing the textual description of the bug reports. A feature vector was generated through the textual information and then presented to a machine learning classifier.

Some other approaches use a more heuristic way to identify bug reports. For example, Zaman et al. (2011) combined keyword searching and statistical sampling to distinguish between performance bugs and security bugs in Firefox bug reports. Gegick et al. (2010) proposed a technique to identify security bug reports based on keyword mining and performed an empirical study based on an industry bug repository.

While all the above work significantly advanced the state-of-the-art, but results related to data mining on software security issues are often problematic:

- Neuhaus et al. (2007) explored the dependency structure within RedHat Linux to learn vulnerability predictors with precision and recall of 83% and 65%. Neuhaus and Zimmermann (2009) later applied their dependency-based methods to the same code base, but at a much larger scale of granularity (system, not specific applications). Their results were not impressive: precision and recall of 40% and 20%, respectively.
- Nguyen and Tran (2010), similarly, applied explored dependency structure. Though not as impressive as Neuhaus and Zimmermann, they achieved precision and recall of 60%

and 61%. However, their code dependency network analysis is not a general method for building vulnerability predictors.

– Scandariato et al. (2014) used a text mining approach over the source code for their vulnerability predictors. They report prediction models with precision and recall over 95%. However, these results were based on a somewhat contentious methodology. The unfiltered alerts of a static code analysis tool were used to label code components as "vulnerable" or not. Such static code analysis tools have a notoriously large false positive rate, declaring that many code components are "vulnerable" when the vulnerabilities are actually false positives.

2.2 FARSEC: Extending Data Mining for Bug Reports

The previous section reported certain problems with existing methods where data mining was applied to security related tasks. In the recently proposed FARSEC (Peters et al. 2018) research, Peters et al. reported more success after focusing on a particular problem within the security domain.

FARSEC is a technique that adds an irrelevancy pruning step to data mining in building security bug reports prediction models. Table 2 lists the filters explored in the FARSEC research. The purpose of filtering in FARSEC is to remove non-security bug reports with security related keywords. To achieve this goal, FARSEC applied an algorithm that firstly calculated the probability of the keywords appearing in security bug report and non-security bug report, and then calculated the score of the keywords.

Inspired by previous works (Graham 2004; Jalali et al. 2008), several tricks were also introduced in FARSEC to reduce false positives. For example, FARSEC built the *farsectwo* filter by multiplying the frequency of non-security bug reports by two, aiming to achieve a good bias. The *farsecsq* filter was created by squaring the numerator of the support function to improve heuristic ranking of low frequency evidence.

In addition, FARSEC also tested a noise detection algorithm called CLNI (Closet List Noise Identification) (Kim et al. 2011). Specifically, CLNI works as follows: During each iteration, for each instance i, a list of closest instances are calculated and sorted according to Euclidean Distance to instance i. The percentage of top N instances with different class values is recorded. If percentage value is larger or equal to a threshold, then instance i is highly probable to be a noisy instance and thus included to noise set S. This process is repeated until two noise sets S_i and S_{i-1} have the similarity over ϵ (e.g., ϵ is 0.99). A threshold score (e.g., 0.75) is set to remove any non-buggy reports above the score.

Table 2 Different filters used in FARSEC

Filter	Description
farsecsq	Apply the Jalali et al. (2008) support function to the frequency of words found in SBRs
farsectwo	Apply the Graham version (Graham 2004) of multiplying the frequency by two.
farsec	Apply no support function.
clni	Apply CLNI filter to non-filtered data.
clnifarsec	Apply CLNI filter to farsec filtered data.
clnifarsecsq	Apply CLNI filter to farsecsq filtered data.
clnifarsectwo	Apply CLNI filter to farsectwo filtered data.

One of the common issues with imbalanced data prediction is the large number of false positives in the prediction results. This matters because it means potentially extra effort is required from developers to check those false positives. FARSEC tries to address this problem by generating a list of ranked bug reports. This method takes two steps. In the first step, for a filter f, the ranked prediction results are selected from non-filtered data or data with filters other than f which has less number of predicted security bug reports than filter f. If the first step does not apply, the chronological order is used in step two. As a result, the predicted security bug reports are close to the top of the list than non-security bug reports.

2.3 Hyperparameter Optimization for Learner and Pre-Processor Options

One data mining approach not fully explored by FARSEC (or much of other works reviewed above) is hyperparameter optimization, i.e. the process of searching the most optimal hyperparameters in data mining learners (Biedenkapp et al. 2018). In machine learning, hyperparameters reflect policies within a model. For example:

- For random forest, a hyperparameter could be the number of trees in the forest.
- For nearest neighbor algorithm, a hyperparameter could be the number of k nearest neighbors used for classification (Keller et al. 1985).
- For text mining, a hyperparameter might control how many words are selected via term weighting.

In this list, the first two are examples of learner hyperparameters while the third one is an example of pre-processor hyperparameter that is selected before the learner executes. Table 1 lists the learner and pre-processor options we explore in this study. The search space of these parameters is shown in Table 3. In those tables, we use the same five machine learning learners as seen in the FARSEC study, i.e., Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Multilayer Perceptron (MLP) and K Nearest Neighbor (KNN). They are widely used for software engineering classification problems (Lessmann et al. 2008). As for the pre-processors, as mentioned in the introduction section, we do not claim that this is the entire set of possible pre-processors. Rather, we just say that any reader of the recent SE data mining literature might have seen many of these. Hence, they might be tempted to try them.

Furthermore, Table 4 shows how often these kinds of hyperparameters have been explored in the previous security relevant literature. As seen from the table:

- A minority of papers have explored learner hyperparameter optimization.
- Only a handful of them have tried pre-processor hyperparameter optimization.
- We have only found one prior work that tried our *dual optimization* approach that explored both pre-processor and learner optimization (Agrawal et al. 2019). However, note that that paper was not in the security domain.

There are good reasons to try and avoid dual optimization – an exhaustive search through all options is computationally intractable. Given N choices for P learner parameters, the space of possible hyperparameter optimizations in $(N)^P$. Worse still, if the space of options increases to include learners and N choices for M pre-processors (such as those listed in Table 1), then the search space is now $(N)^{P+M}$, i.e. exponentially larger.

It is neither useful nor practical to explore such a large space of options via exhaustive search. For example, *grid search* (Bergstra et al. 2011; Tantithamthavorn et al. 2016) is a "brute force" hyperparameter optimizer that wraps a learner into for-loops that walk

Table 3 List of hyperparameters optimized in different learners and pre-processors. The brief description of each learner and pre-processor can be found in Table 1

Type	Name	Parameters	Default	Tuning Range
Learner	Random Forest	n_estimators	10	[10, 150]
		min_samples_leaf	1	[1, 20]
		min_samples_split	2	[2, 20]
		max_leaf_nodes	None	[2, 50]
		max_features	auto	[0.01, 1]
		max_depth	None	[1, 10]
	Logistic Regression	C	1.0	[1.0, 10.0]
		max_iter	100	[50, 200]
		verbose	0	[0, 10]
	Multilayer Perceptron	alpha	0.0001	[0.0001, 0.001]
		learning_rate_init	0.001	[0.001, 0.01]
		power_t	0.5	[0.1, 1]
		max_iter	200	[50, 300]
		momentum	0.9	[0.1, 1]
		n_iter_no_change	10	[1, 100]
	K Nearest Neighbor	leaf_size	30	[10, 100]
		n_neighbors	5	[1, 10]
	Naive Bayes	var_smoothing	1e-9	[0.0, 1.0]
Pre-processor	SMOTE	k	5	[1, 20]
		m	50 %	[50, 400]
		r	2	[1, 6]
	Normalizer	norm	l2	[l1, l2, max]
		copy	True	[True, False]
	StandardScaler	copy	True	[True, False]
		with_mean	True	[True, False]
		with_std	True	[True, False]
	MinMaxScaler	copy	True	[True, False]
		min	0	[-5, 0]
		max	1	[1, 5]
	MaxAbsScaler	copy	True	[True, False]
	RobustScaler	with_centering	True	[True, False]
		with_scaling	True	[True, False]
		q_min	25.0	[10, 40]
		q_max	75.0	[60, 90]
		copy	True	[True, False]
	QuantileTransformer	n_quantiles	1000	[10, 2000]
		output_distribution	uniform	[uniform, normal]
		ignore_implicit_zeros	False	[True, False]
		subsample	1e5	[100, 150000]
		copy	True	[True, False]

Table 3 (continued)

Type	Name	Parameters	Default	Tuning Range
	PowerTransformer	method	yeo-johnson	[yeo-johnson, box-cox]
		standardize	True	[True, False]
		copy	True	[True, False]
	Binarization	threshold	0.0	[0, 10]
		copy	True	[True, False]
	PolynomialFeatures	degree	2	[2, 4]
		interaction_only	False	[True, False]
		include_bias	True	[True, False]
		order	C	[C, F]

through a wide range of all learner's control parameters. Simple to implement, it has many drawbacks. Firstly, even this brute force approach does not sample all the options since its for-loops jump over numeric ranges using some increment value. This means that grid search can actually skip over the important optimizations. Secondly, it suffers from the "curse of dimensionality". That is, after just a handful of options, grid search can miss important optimizations. Thirdly, and worse still, much CPU resources can be wasted during grid search since experience has shown that only a few ranges within a few optimization parameters really matter (Bergstra and Bengio 2012).

An alternative to grid search is the *random search* (Bergstra and Bengio 2012) that stochastically samples the search space and evaluates sets from a specified probability distribution. Evolutionary algorithms are a variant of random search that runs in "generations" where each new generation is seeded from the best examples selected from the last generation (Goldberg 2006). Simulated annealing is a special form of evolutionary algorithms where the population size is one (Kirkpatrick et al. 1983; Menzies et al. 2007b).

Genetic algorithms (GA) is another form of random search where the population size is greater than one, and new mutants are created by crossing over parts of the better members of the current population (Goldberg 2006; Panichella et al. 2013). Note one feature of genetic algorithms is that, their mutation operator never changes during the execution of the GA. That is, GAs have no facility for using experience from the domain to define better mutators.

Another kind of random search, that does use domain experience to define better mutators, is *differential evolution* (DE) (Storn and Price 1997). In differential evolution algorithm, the size of a mutation is selected from a pool of previous cache of "superior" mutations; i.e. mutants that are known to be better than other mutants. That is, as differential evolution algorithm learns more and more about what mutants are superior, it is also learning how better to mutate old individuals into better ones. There are four major steps in differential evolution algorithm – *initialization*, *mutation*, *crossover*, and *selection*:

- The *initialization* step creates a population of individuals, while each individual is an instance of the parameters generated randomly within given bounds.
- In the *mutation* step, for each individual p_i in the population, three other individuals a, b, c (not the current one) are randomly selected. A mutant individual is created by combining these three selected individuals. The difference is then computed between

Table 4 List of previous research studies that address security and software engineering problems

Reference	Year	Citation	Learner optimization	Pre-processor optimization	Security related
(Thornton et al. 2013)	2013	754	✓	×	×
(Li et al. 2017)	2017	358	✓	×	×
(Lamkanfi et al. 2010)	2010	285	×	×	✓
(Sun et al. 2011)	2011	264	×	×	×
(Feurer et al. 2015)	2015	193	✓	×	×
(Gegick et al. 2010)	2010	146	×	×	✓
(Xia et al. 2017)	2017	139	✓	×	×
(Tian et al. 2012)	2012	133	✓	×	✓
(Fu et al. 2016)	2016	100	✓	×	×
(Tian et al. 2015)	2015	64	×	×	✓
(Wang and Xu 2018)	2018	60	✓	×	×
(Agrawal et al. 2018)	2018	59	×	✓	✓
(Lazar et al. 2014)	2014	54	×	×	×
(Agrawal and Menzies 2018)	2018	49	×	✓	×
(Xia et al. 2014)	2014	44	×	×	×
(Tantithamthavorn et al. 2018)	2018	34	×	✓	×
(Hindle et al. 2016)	2016	29	×	×	×
(Nair et al. 2018)	2018	29	✓	×	×
(Zhang et al. 2015)	2015	28	×	×	✓
(Wijayasekara et al. 2014)	2014	26	×	×	✓
(Yang et al. 2017)	2017	23	×	×	✓
(Chan et al. 2013)	2013	20	✓	×	×
(Di Francescomarino et al. 2018)	2018	20	✓	×	×
(Deshmukh et al. 2017)	2017	18	✓	×	×
(Xia et al. 2016)	2016	17	×	×	×
(Osman et al. 2017)	2017	14	✓	×	✓
(Menzies et al. 2018)	2018	16	✓	×	×
(Yang et al. 2016)	2016	11	×	×	✓
(Goseva-Popstojanova and Tyo 2018)	2018	9	×	×	✓
(Han et al. 2018)	2018	6	×	×	×
(Agrawal et al. 2019)	2019	4	✓	✓	×

In this list, only one prior publication optimized both the learner and pre-processor (see the last line) and that paper did not explore the security domain. This list of papers was found either from the above literature review or from Google Scholar using the search query, e.g., *"((hyperparameter optimization) and (security)) or ((hyperparameter optimization) and (security bug reports)) or (optimization and security) or (optimization and (pre-processors) and security), ((hyperparameter optimization) and (software engineering))"*. These queries returned more than 5,000 papers which were further pruned. We only used papers in the last ten years (2010-2020) and which had appeared in (a) top conferences or (b) venues listed by Google Scholar as "top-ranked" (e.g., see tiny.cc/top20soft_venues)

two individuals and added to the rest individual after multiplying a mutation factor to the difference, i.e., $y_k = a_k + f \times (b_k - c_k)$. The mutation factor f is a positive number that controls the amplification difference between two individuals.

- At some crossover probability cf, the mutant attribute is then added to a vector that is the new mutant in the $crossover$ step.
- Finally, during the *selection* step, differential evolution algorithm decides if the mutant generated from a, b, c is better than p_i. If so, the mutant replaces p_i and the algorithm moves on to some other member of the population p_j.
- All the above steps have to be repeated again for the remaining individuals p_j, which completes the first iteration of the algorithm. After this process, some of the original individuals of the population will be replaced by better ones. That is, all subsequent mutants will be built from the "superior" examples cached in the population.

As to the control parameters of the differential evolution algorithm, using advice from the differential evolution algorithm user group (see tiny.cc/how2de), we set $\{np, f, cr\} = \{10k, 0.8, 0.9\}$, where k is the number of parameters to optimize, and np is the size of whole population. Note that we set the number of iteration $\{g\}$ to 3, 10, which are denoted as DE3 and DE10 respectively. A small number (i.e., 3) is used to test the effects of a CPU-light effort estimator. A larger number (i.e., 10) is selected to check if anything is lost by restricting the inference to small iterations.

In the software engineering literature, differential evolution algorithm has been seen to outperform other methods such as (a) particle swarm optimization (Vesterstrøm and Thomsen 2004); (b) the grid search used by Tantithamthavorn et al. (2016) to optimize their defect predictors; or (c) the genetic algorithm used by Panichella et al. (2013) to optimize a text miner. Also, the differential evolution algorithm has been proven useful in prior software engineering optimization studies (Fu et al. 2016).

3 SWIFT: The Dual Optimization Approach

Recent studies show substantial interest in automated hyperparameter optimization on complex and computational expensive machine learning models with many hyperparameters. By tailoring the models to the problems at hand, hyperparameter optimization improves the model performance and even leads to new state-of-the-art results.

Apart from machine learning models, data pre-processing techniques are often involved in practical machine learning pipeline. Real-world data is often inconsistent, lacking in certain behaviors of trends, or even contains many errors. Data pre-processing transforms the raw data into a more useful and efficient shape. Similar to model optimization, pre-processing optimization also shows increasing interest (Agrawal and Menzies 2018).

While each individual optimization problem already experiences computational complexity, for example, Tables 1 and 3 demonstrate a list of machine learning learners and data pre-processing techniques, as well as their hyperparameter options. Even this partial list includes thousands of configuration options. The cost of running an optimizer through these options would be quite expensive, requiring days to weeks of CPU resources (Tantithamthavorn et al. 2016, 2018). A combination of the above two optimization problems (i.e., dual optimization) faces even more challenges.

A "simpler" optimizer is required to tackle the dual optimization challenge. This ideal optimizer should be able to achieve better performance than each individual optimizer and the computational complexity would not increase.

In 2005, Deb et al. (2005) proposed an idea named ϵ-*dominance* that partitions the output space of an optimizer into ϵ-sized grids. The principle of this idea is that if there exists

some ϵ value below which it is useless or impossible to distinguish the results, then it superfluous to explore anything less than ϵ. Specifically, consider the bug reports classification task discussed in this paper, if the performances of two learners (or a learner with various hyperparameters) differ in less than some ϵ value, then we cannot statistically distinguish them. For the learners which do not significantly improve the performance, we can further reduce the attention on them.

Inspired by the idea of ϵ-dominance, we propose a method named SWIFT to address the dual optimization problem. From a high level, SWIFT is essentially a tabu search; i.e., if some settings resulted in some performance within ϵ of any older result, then SWIFT marked that option as "to be avoided". SWIFT applies "item ranking" in seeking optimal learner and pre-processor, and further refines their option ranges. SWIFT returned the best setting seen during the following three stage process:

- *Initialization*: all option items i are assigned equal weightings.
- The *item ranking* stage reweights items i in column 2 of Table 3; e.g. terms like "Random Forest" or "RobustScaler".
- The *numeric refinement* stage adjusts the tuning ranges of the last column in Table 3.

In summary, what is happening here is that item selection handles the "big picture" decisions about what pre-processor or learner to use while numeric refinement focuses on smaller details about numeric ranges.

More specifically, the algorithm runs as follows:

- *Initialization*: Assign weights $w_i = 0$ to all items i in column 2 of Table 3.
- *Item ranking*: N_1 times, we make a random selection of a learner and pre-processor from column 2, favoring those items with higher weights. For the selected items, we select a value at random from the "Tuning Range"s of the last column of Table 3. Using that selection, we build a model and evaluate it on test data. If we obtain a model whose performance is more/less than ϵ of any prior results, then we add/subtract (respectively) 1.0 from w_i.
- *Numeric refinement*: N_2 times, we refine the numeric tuning ranges (lo, hi) seen in the last column of Table 3. In this step, the item ranking continues. But now, if ever some numeric tuning value $lo \leq b \leq hi$ produces a better model, then we adjust that range, as follows. Whichever of $x \in (lo, hi)$ that is the furthest from b is moved to $(b + x)/2$.

(Aside: It should be pointed out that SWIFT is not a multi-objective optimization problem. We choose g-measure as our optimization goal (i.e., the aim to increase). G-measure is the harmonic mean of recall and the complement of false alarms. More description of this metric and the reason of the choice are further discussed in Section 4.4.)

Agrawal et al. (2019) have successfully applied ϵ-dominance to some SE tasks such as software defect prediction and SE text mining, and they proposed the approach named DODGE. For the cases studied by DODGE, that approach was able to explore a large space of hyperparameter options, while at the same time generated models that performed as well or better than the prior state-of-the-art in defect prediction and SE text mining (Agrawal et al. 2019). SWIFT is an improved version of DODGE since we found that DODGE cannot be directly applied to our bug report data without any modification effort. There are several reasons for this after investigation.

Firstly, DODGE guided its optimization using metrics that were alien to this domain. For example, the "Popt20" goal used in the original DODGE studied by Agrawal et al. (2019) optimizes for an economic concern not explored by Peters et al. in the FARSEC study.

Popt20 is relevant to general SE tasks, but not for security-related domains. Specifically, we want to find as many of the security bug reports as possible, even if that means developers have to spend some time exploring a few more false positives. Accordingly, we swapped out Popt20 in favor of the "g-measure" as defined in Section 4.

Second, once we changed evaluation goals, another concern became apparent. We found that the distribution of the w_i weights was far more skewed in the security bug report data than in the other kinds of software engineering tasks studied by Agrawal et al. This skewed data meant that, usually, there was only one good learner and one good data pre-processor for the security data sets. We conjecture that this is so since we require specific biases to find the target concept of something so particular as a security bug report. For the original version of DODGE, such skewed w_i weights are a problem since, as mentioned above, item ranking continues during the numeric refinement stage.

SWIFT is specifically designed for our security data. SWIFT is designed to make better use of the w_i skews. After item ranking, SWIFT only takes the best learner and data pre-processor forward into numeric refinement. While the above two changes were only a small coding change to the original DODGE, their effects were profound.

4 Experiment

4.1 Hyperparameter Optimization Ranges

This paper compares SWIFT against the differential evolution algorithm (described in Section 2) since recent papers at ICSE (Agrawal and Menzies 2018) the IST journal (Agrawal et al. 2018) reported that the differential evolution algorithm can find large improvement in learner performance for SE data. Table 5 lists the control settings for the differential evolution algorithm used in this paper (that table was generated by combining the advice at the end of Section 2.3 with Table 3). For SWIFT, we used the settings recommended by Agrawal et al. (2019). Note that proving the optimum of our solution is not the goal of this paper. In fact, like Wolpert (Wolpert and Macready 1997), we doubt if there is any "best" optimizer that works for all data (for more on that, see the "No Free Lunch" theorem discussion (Wolpert and Macready 1997) in search and optimization). Therefore, this paper is not searching for the "best" result, but rather it is searching for "better" than the prior state-of-the-art.

Note that SWIFT and differential evolution algorithm were applied to learners from the scikit-learn toolkit (Pedregosa et al. 2011). Table 3 lists all the hyperparameters we select for both data mining learners and data pre-processors based on scikit-learn.

We choose not to explore other hyperparameter optimizers, for pragmatic reasons. Numerous other studies have shown that the differential evolutionary algorithm (DE) well performed for optimization problems (Menzies et al. 2018; Fu and Menzies 2017; Fu et al. 2016; Wang et al. 2015; Yildizdan and Baykan 2020; Onan et al. 2016). If our goal was to claim that DE was somehow the optimal optimizer, we would have to perform a wider range of study of optimizers (i.e more than just DE). However, our goal is not that (and, in fact, there are support theoretical reasons for assuming that no optimizer is ever "best" for all data sets (Wolpert and Macready 1997)). Rather, our purpose is to provide an improvement on the prior state-of-the-art (the FARSEC paper). As shown below, that can be achieved using DE. While in future work we aim to explore other optimizers, for the purposes of this paper, using DE is enough.

4.2 Data

For this work, we compare the differential evolutionary algorithm (DE) and SWIFT to FARSEC using the same data as used in the FARSEC study. The data set includes five projects: four from Apache projects (i.e., Ambari, Camel, Derby and Wicket) (Ohira et al. 2015) and one from the Chromium project. For the Apache projects, one thousand bug reports are randomly selected for each project with BUG or IMPROVEMENT label from the JIRA bug tracking system (Ohira et al. 2015). All the selected bug reports are then classified with scripts or manually into six high impact bugs (i.e., *Surprise*, *Dormant*, *Blocking*, *Security*, *Performance*, and *Breakage* bugs). All the target bug reports in our data set all belong to *Security* bug reports (i.e., bug reports of the type *Security*). For the Chromium project, security bugs are labeled as *Bug-Security* when submitted to bug tracking systems. All other types of bug reports in the data set are treated as non-security bug reports.

The datasets from FARSEC are publicly available. Our experiments reproduce and improve the FARSEC results using the same datasets. Table 6 shows the characteristics of the FARSEC datasets. As we see from the table, one unique feature of the data set is the rarity of the target class. The "SBRs %" column in both training and testing data set indicates that security bug reports make up a very small percentage of the total number of bug reports in projects like Chromium.

4.3 Experimental Rig

Our experiment design is mainly divided into two parts. When we optimize learners or data pre-processors individually, we divide each *training data* into $B = 10$ bins, and validate our models using bin B_i after training them on *training data* - B_i. This 10-fold cross-validation is used to pick the best candidate learner/pre-processor as well as their hyperparameters with the highest performance for that data set. We also need to point out that the 10-fold cross-validation does not apply to the dual optimization, and the way we select the best candidate learner and pre-processor in SWIFT is based on weight calculation and we further refine their hyperparameter's numeric ranges as we discuss in Section 3.

After finding the best learners and/or pre-processors, we then train the models with the whole training dataset, and test on the separate testing dataset as FARSEC.

Table 5 List of parameters in differential evolution (DE) algorithm for different learners and pre-processor

Learner & Pre-processor	DE Parameter			
	NP	F	CR	ITER
Random Forest	60	0.8	0.9	3, 10
Logistic Regression	30			
Multilayer Perceptron	60			
K Nearest Neighbor	20			
Naive Bayes	10			
SMOTE	30	0.8	0.9	10

* Note: **NP** is the size of population; **F** is the parameter controlling the differential weight; **CR** is the probability threshold; **ITER** is the number of iterations.

Table 6 Imbalanced characteristic of bug report data sets from FARSEC (Peters et al. 2018)

		Training			Testing		
Project	Filter	#SBRs	#BRs	SBRs(%)	#SBRs	#BRs	SBRs(%)
Chromium	train	77	20,970	0.37	115	20,970	0.55
	farsecsq		14,219	0.54			
	farsectwo		20,968	0.37			
	farsec		20,969	0.37			
	clni		20,154	0.38			
	clnifarsecsq		13,705	0.56			
	clnifarsectwo		20,152	0.38			
	clnifarsec		20,153	0.38			
Wicket	train	4	500	0.80	6	500	1.20
	farsecsq		136	2.94			
	farsectwo		143	2.80			
	farsec		302	1.32			
	clni		392	1.02			
	clnifarsecsq		46	8.70			
	clnifarsectwo		49	8.16			
	clnifarsec		196	2.04			
Ambari	train	22	500	4.40	7	500	1.40
	farsecsq		149	14.77			
	farsectwo		260	8.46			
	farsec		462	4.76			
	clni		409	5.38			
	clnifarsecsq		76	28.95			
	clnifarsectwo		181	12.15			
	clnifarsec		376	5.85			
Camel	train	14	500	2.80	18	500	3.60
	farsecsq		116	12.07			
	farsectwo		203	6.90			
	farsec		470	2.98			
	clni		440	3.18			
	clnifarsecsq		71	19.72			
	clnifarsectwo		151	9.27			
	clnifarsec		410	3.41			
Derby	train	46	500	9.20	42	500	8.40
	farsecsq		57	80.70			
	farsectwo		185	24.86			
	farsec		489	9.41			
	clni		446	10.31			
	clnifarsecsq		48	95.83			
	clnifarsectwo		168	27.38			
	clnifarsec		435	10.57			

4.4 Evaluation Metrics

To understand the open issues with bug report classification, firstly we must define how they are **assessed**. If (TN, FN, FP, TP) are the true negatives, false negatives, false positives, and true positives, respectively, found by a classifier, then:

- *pd* = Recall = TP/(TP+FN), the percentage of the actual security bug reports that are predicted to be security bug reports.
- *pf* = False Alarms = FP/(FP+TN), the percentage of the non-security bug reports that are reported as security bug reports.
- *prec* = Precision = TP/(TP+FP), the percentage of the predicted security bug reports that are actual security bug reports.
- *f-score* = F-Measure = 2*pd*prec/(pd+prec), the harmonic mean of the model's precision and recall.

This paper adopts the same evaluation criteria as the original FARSEC paper; i.e. the recall (*pd*) and false alarm (*pf*) measures. Also, to control the optimization algorithm, we are endeavoring to minimize false alarms while maximizing recall. To achieve those goals, we maximize the *g-measure* which is the harmonic mean of recall and the complement of false alarms in our algorithm.

$$g = \frac{2 \times pd \times (1 - pf)}{pd + (1 - pf)} \tag{1}$$

g is maximal when *both* recall (*pd*) is high and false alarm (*pf*) is low.

We choose *g-measure* based on the following considerations. For an imbalanced dataset where there is a skew in the class distribution (e.g., negative samples are much more than positive samples), we have two competing goals:

- On the one hand, we want to focus on minimizing false negatives (i.e., security bug reports are not missed in prediction (Scandariato et al. 2014)).
- On the other hand, we prefer not to predict too many non-security bug reports as security bug reports, which is $(1 - pf)$ that also represents specificity.

As to why we use these measures but not some others such as precision, Menzies et al. (2007a) argue that when the target class is less than 10% (as is with all our data), the precision results become more a function of the random number generator used to divide data (for testing purposes). Therefore, we cannot recommend precision for this kind of data. (Aside: we are not alone in this view (that precision should not be used). For example, the FARSEC paper (that this work builds on) did not assess its models via precision.)

Besides the above, we also use another evaluation measure called IFA (Initial False Alarm) to evaluate the performance. IFA is the number of initial false alarm encountered before we make the first correct prediction (Huang et al. 2017, 2019). IFA is widely used in defect prediction, and previous works (Kochhar et al. 2016; Parnin and Orso 2011) have shown that developers are not willing to use a prediction model if the first few recommendations are all false alarms.

Furthermore, metrics like recall and g-measure are set-based measures, and they are computed using unordered sets of data. To evaluate the results of ranking bug report, mean average precision (MAP) is commonly used to indicate the quality of a ranking by comparing with the ground truth. A higher MAP value usually means more actual security bug reports that predicted are close to the top of the list.

Equations (2) and (3) show how average precision (AP) and MAP are computed. Specifically, AP_n is the average of precision @k where $P(k)$ is the precision at point k in the ranked list and n is the number of predicted security bug reports. As done in the FARSEC paper, we say that MAP_n is the mean of cumulative average precision scores for each decile.

$$AP_n = \sum_{k=1}^{n} \frac{P(k)}{n} \tag{2}$$

$$MAP_n = \sum_{i=1}^{N} \frac{AP_{ni}}{N} \tag{3}$$

4.5 Statistics

This study ranks treatments using the Scott-Knott procedure recommended by Mittas & Angelis in their 2013 IEEE TSE paper (Mittas and Angelis 2013). This method sorts results from different treatments, then splits them in order to maximize the expected value of differences in the observed performances before and after divisions. For lists l, m, n of size ls, ms, ns where $l = m \cup n$, the "best" division maximizes $E(\Delta)$; i.e. the difference in the expected mean value before and after the spit:

$$E(\Delta) = \frac{ms}{ls} abs(m.\mu - l.\mu)^2 + \frac{ns}{ls} abs(n.\mu - l.\mu)^2$$

Scott-Knott then checks if that "best" division is actually useful. To implement that check, Scott-Knott would apply some statistical hypothesis test H to check if m, n are significantly different (and if so, Scott-Knott then recurses on each half of the "best" division). For this study, our hypothesis test H was a conjunction of the A12 effect size test of and non-parametric bootstrap sampling; i.e. our Scott-Knott divided the data if *both* bootstrapping and an effect size test agreed that the division was statistically significant (95% confidence) and not a "small" effect ($A12 \geq 0.6$).

For a justification of the use of non-parametric bootstrapping, see Efron & Tibshirani (1994, p220–223). For a justification of the use of effect size tests see Kampenes et al. (2007) who warn that even if a hypothesis test declares two populations to be "significantly" different, then that result is misleading if the "effect size" is very small. Hence, to assess the performance differences we first must rule out small effects. Vargha and Delaney's non-parametric A12 effect size test explores two lists M and N of size m and n:

$$A12 = \left(\sum_{x \in M, y \in N} \begin{cases} 1 & if\ x > y \\ 0.5 & if\ x == y \end{cases} \right) / (mn)$$

This expression computes the probability that the numbers in one sample are bigger than in another. This test was endorsed by Arcuri and Briand (2011). Tables 7, 8 and 9 present the results of our Scott-Knott procedure for each project data set. These results are discussed, extensively, in the next section.

5 Results

In this section, Tables 7, 8 and 9 report results with and without hyperparameter optimization of the pre-processors or learners or both. For the sake of completeness, we also add

Table 7 **RQ1** results: recall

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
Chromium	train	15.7	46.9	68.7	73.9	**86.1**
	farsecsq	14.8	64.3	80.0	**84.3**	72.2
	farsectwo	15.7	40.9	**78.3**	77.4	77.4
	farsec	15.7	46.1	**80.8**	72.2	77.4
	clni	15.7	30.4	74.8	72.2	**80.9**
	clnifarsecsq	49.6	72.2	82.6	**86.1**	72.2
	clnifarsectwo	15.7	50.4	**79.1**	74.8	78.3
	clnifarsec	15.7	47.8	**78.3**	74.7	72.2
	Median Recall	15.7	47.3	78.7	74.8	77.4
Wicket	train	16.7	0.0	**66.7**	**66.7**	50.0
	farsecsq	66.7	50.0	**83.3**	**83.3**	**83.3**
	farsectwo	**66.7**	50.0	**66.7**	**66.7**	**66.7**
	farsec	33.3	**66.7**	**66.7**	**66.7**	**66.7**
	clni	0.0	16.7	**50.0**	**50.0**	**50.0**
	clnifarsecsq	33.3	**83.3**	**83.3**	**83.3**	**83.3**
	clnifarsectwo	33.3	50.0	**66.7**	**66.7**	**66.7**
	clnifarsec	50.0	**66.7**	**66.7**	**66.7**	**66.7**
	Median Recall	33.3	50.0	66.7	66.7	66.7
Ambari	train	14.3	28.6	57.1	57.1	**85.7**
	farsecsq	42.9	57.1	57.1	57.1	**85.7**
	farsectwo	57.1	57.1	57.1	57.1	**85.7**
	farsec	14.3	57.1	57.1	57.1	**85.7**
	clni	14.3	28.6	57.1	57.1	**85.7**
	clnifarsecsq	57.1	57.1	57.1	57.1	**71.4**
	clnifarsectwo	28.6	57.1	57.1	57.1	**85.7**
	clnifarsec	14.3	57.1	57.1	57.1	**85.7**
	Median Recall	21.5	57.1	57.1	57.1	85.7
Camel	train	11.1	16.7	33.3	44.4	**55.6**
	farsecsq	16.7	44.4	44.4	55.6	**66.7**
	farsectwo	50.0	44.4	**61.1**	**61.1**	**61.1**
	farsec	16.7	22.2	33.3	33.3	**55.6**
	clni	16.7	16.7	33.3	38.9	**50.0**
	clnifarsecsq	16.7	38.9	27.8	33.3	**61.1**
	clnifarsectwo	11.1	61.1	**72.2**	61.1	61.1
	clnifarsec	16.7	22.2	33.3	38.9	**55.6**
	Median Recall	16.7	38.5	33.3	41.7	58.4
Derby	train	38.1	47.6	54.7	59.5	**69.0**
	farsecsq	54.8	59.5	54.7	**66.7**	**66.7**

Table 7 (continued)

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
	farsectwo	47.6	59.5	47.6	66.7	**78.6**
	farsec	38.1	47.6	57.1	59.5	**64.3**
	clni	23.8	45.2	57.7	61.9	**69.0**
	clnifarsecsq	54.8	59.5	76.2	**69.0**	66.7
	clnifarsectwo	35.7	59.5	54.8	61.9	**66.7**
	clnifarsec	38.1	47.6	61.9	57.1	**66.7**
	Median Recall	38.1	53.6	56.0	61.9	66.7
Overall Median Recall		21.5	50.0	57.1	61.9	66.7

In these results, *higher* recalls (a.k.a. pd) are *better*. For each row, the best results are highlighted in boldface (these are the cells that are statistically the same as the best median result – as judged by our Scott-Knot test). Across all rows, SWIFT has the most number of best results

results of precision and f-measure in Tables 10 and 11. Using those results, we can now answer our proposed research questions.

5.1 RQ1

RQ1. Can hyperparameter optimization techniques improve the performance of models that better distinguish security bug reports from other bug reports?

5.1.1 Recall Results

In the recall results of Table 7, we can observe that FARSEC rarely achieves the best results while SWIFT is much better than FARSEC. For example:

- In the Chromium project, median recall changes from 15.7% to 77.4% from FARSEC to SWIFT.
- In the Ambari project, the median recall changes from 21.5% to 85.7% from FARSEC to SWIFT.
- Overall, as shown in the last line of Table 7, the improvement is from 21.5% to 66.7% (FARSEC to SWIFT).

In addition, in Table 7, the boldface cells show the "best" results in each row (where "best" is defined using the statistical significance tests of Section 4.5). Overall, SWIFT is statistically significantly best in 31/40 of all the rows of Table 7. This is more than twice as many wins as other approaches explored in this table; e.g. DE+pre-processors scores best in only 13/40 rows. Hence, for this data set, we say that dual optimization of both learners and pre-processors work best.

Just for completeness, we note that for all methods with any data pre-processing procedure (i.e., in the last three columns of Table 7) work well for the Wicket project. Clearly, for

Table 8 **RQ1** results: false positive rate (a.k.a., pf), the lower values are better

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
Chromium	train	**0.2**	6.8	24.1	17.8	24.0
	farsecsq	**0.3**	10.3	31.5	25.1	14.3
	farsectwo	**0.2**	6.5	27.6	23.1	26.1
	farsec	**0.2**	6.9	36.1	14.9	14.7
	clni	**0.2**	4.1	24.8	13.6	26.2
	clnifarsecsq	**3.8**	14.2	30.4	25.6	14.0
	clnifarsectwo	**0.2**	7.0	29.9	12.8	18.9
	clnifarsec	**0.2**	10.4	29.0	17.1	20.2
	Median FPR	0.2	7.0	29.5	17.5	19.5
Wicket	train	7.1	**5.1**	32.0	12.1	27.5
	farsecsq	**38.3**	44.5	71.3	66.8	66.7
	farsectwo	**36.6**	42.3	68.2	62.9	61.5
	farsec	**8.1**	23.1	43.9	26.1	23.3
	clni	5.5	**2.4**	21.1	12.5	14.4
	clnifarsecsq	**25.5**	66.8	66.8	66.8	57.5
	clnifarsectwo	**27.7**	39.9	61.3	61.3	52.8
	clnifarsec	**10.5**	23.1	38.9	22.9	22.1
	Median FPR	18.0	31.5	52.6	43.7	40.2
Ambari	train	1.6	**0.8**	20.1	10.8	17.8
	farsecsq	14.4	**2.8**	30.4	17.2	23.7
	farsectwo	3.0	**2.8**	22.1	17.8	19.7
	farsec	4.9	**2.0**	19.9	7.1	20.3
	clni	2.6	**0.8**	12.4	8.9	18.1
	clnifarsecsq	7.7	**2.4**	13.4	7.1	29.0
	clnifarsectwo	4.5	**2.8**	13.0	5.1	22.7
	clnifarsec	**0.0**	2.4	7.9	3.9	18.9
	Median FPR	3.8	2.4	16.7	8.0	20.0
Camel	train	3.5	**1.5**	27.4	35.9	15.8
	farsecsq	**11.4**	24.7	20.5	23.4	27.8
	farsectwo	41.8	**17.6**	71.0	53.1	45.2
	farsec	**6.9**	12.4	39.4	28.0	35.7
	clni	12.3	**7.9**	33.6	35.3	24.7
	clnifarsecsq	13.9	14.9	**12.4**	15.6	27.2
	clnifarsectwo	**7.7**	50.0	64.9	51.9	38.8
	clnifarsec	**5.0**	11.6	24.9	34.4	37.1
	Median FPR	9.6	13.7	30.5	34.8	31.8
Derby	train	**6.8**	39.3	22.2	20.7	19.7
	farsecsq	29.9	40.6	51.7	51.5	**22.5**

Table 8 (continued)

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
	farsectwo	**12.4**	24.2	27.9	33.6	40.0
	farsec	6.3	**4.1**	21.0	19.0	13.8
	clni	**0.4**	3.5	16.8	24.5	25.5
	clnifarsecsq	**29.9**	42.4	74.7	65.1	42.3
	clnifarsectwo	**9.2**	24.2	36.5	30.3	52.2
	clnifarsec	6.8	**3.9**	28.8	10.9	19.6
	Median FPR	8.0	24.2	28.4	27.4	24.0
Overall Median FPR		8.0	13.7	29.5	27.4	24.0

Same as Table 7; i.e. the best results are highlighted in boldface. While FARSEC has the most best results, these low false positive rates are only achieved by settling for low recalls (see Table 7)

this data set, data pre-processing such as repairing the class imbalance issue is essential for good performance.

5.1.2 False Positive Rate Results

As to the false positive rate results, Table 8 shows that FARSEC has the lowest false positive rate across more than half of the datasets with filters. However, as shown in Table 7, FARSEC achieves those low false positive rate by settling for some low recalls.

As to SWIFT, we note that its improvements in recall (seen above) come at the cost of some increments in false positive rate. As shown in the last line of Table 8, the overall median false positive rate increases from 8% to 24% (FARSEC to SWIFT). While, ideally, the false positive rate is zero, it is inevitable that there is some cost in dealing with security problems. Another way to look at this is to say while our methods help distinguishing security bug reports (from other bug reports), they also highlight the costs involved in securing software. Our method can better distinguish security bug reports than the prior state-of-the-art. However, to do so, there is some increase in the workload of developers who have to read more code and suffer a (slightly) higher false positive rate. Such is the price of software quality assurance.

Hence we say that this 16% increase in overall false positive rates is the *acceptable* and *inevitable* "price" of increasing recall. As to *acceptable*, the overall false alarms are still less than a quarter – which is in the same range as many other software analytic applications.[1]

As to *inevitable*, consider two models:

– One just predicts "yes" all the time. This model has 100% recall (since it finds every target class) but it suffers from large false positive rates.

[1] e.g. Fig. 12 of Menzies et al. (2007c) lists nine SE data mining applications with median false positive rates of 25%.

Table 9 **RQ1** results: initial false alarm (IFA)

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
Chromium	train	N/A	62	75	61	**58**
	farsecsq	N/A	**20**	72	54	36
	farsectwo	N/A	**37**	91	78	87
	farsec	N/A	62	112	62	**56**
	clni	N/A	**41**	86	48	74
	clnifarsecsq	N/A	41	57	62	**37**
	clnifarsectwo	N/A	**37**	89	47	58
	clnifarsec	N/A	62	113	63	**54**
	Median IFA	N/A	41	88	62	57
Wicket	train	N/A	**25**	60	34	46
	farsecsq	N/A	**29**	37	33	39
	farsectwo	N/A	32	35	34	**31**
	farsec	N/A	23	44	30	**22**
	clni	N/A	**12**	44	21	27
	clnifarsecsq	N/A	9	8	9	**6**
	clnifarsectwo	N/A	**8**	11	12	**8**
	clnifarsec	N/A	17	33	**15**	18
	Median IFA	N/A	20	36	26	25
Ambari	train	N/A	7	8	9	**4**
	farsecsq	N/A	8	21	14	**7**
	farsectwo	N/A	**1**	19	12	3
	farsec	N/A	**1**	35	24	17
	clni	N/A	**1**	32	19	13
	clnifarsecsq	N/A	**8**	18	10	**8**
	clnifarsectwo	N/A	**7**	28	8	11
	clnifarsec	N/A	5	10	**4**	17
	Median IFA	N/A	6	20	11	10
Camel	train	N/A	**6**	19	23	15
	farsecsq	N/A	23	29	32	**14**
	farsectwo	N/A	**4**	13	8	25
	farsec	N/A	17	21	20	**8**
	clni	N/A	**16**	37	33	30
	clnifarsecsq	N/A	5	**3**	**3**	4
	clnifarsectwo	N/A	19	22	15	**12**
	clnifarsec	N/A	**14**	23	29	22
	Median IFA	N/A	15	22	22	15
Derby	train	N/A	4	6	3	**2**
	farsecsq	N/A	**4**	**4**	**4**	**4**

Table 9 (continued)

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
	farsectwo	N/A	4	**3**	5	**3**
	farsec	N/A	**1**	8	7	4
	clni	N/A	**1**	8	5	3
	clnifarsecsq	N/A	**1**	2	2	**1**
	clnifarsectwo	N/A	**2**	9	8	4
	clnifarsec	N/A	**1**	3	3	2
	Median IFA	N/A	2	5	5	3
Overall median IFA		N/A	15	22	22	15

IFA is the number of false alarms developers must suffer through before finding their first target. Lower values are better. Same as format as Table 7; i.e. best results are shown in boldface

– Another model just predicts "no" all the time. This second model has 0% false positive rate (i.e., it never makes mistakes in prediction) but it also has a 0% recall (since it never finds any target class).

In practice, all learners make trade-offs between recall and false positive rate as they explore models somewhere on a curve between:

– Recall from 0% to 100%
– False positive rate from 0% to 100%
– In addition, unless the learner is broken, this curve bends upwards away from the recall == false positive rate line towards the point recall=100% and false positive rate=0% (but rarely does any learner reach this point).

This means that as a learner tries different models, increased recall comes at the cost of also increasing false positive rates. The trick here is to increase recall *more than* false positive rate, as is done by SWIFT. In this paper, we show that we can increase median recall from 21.5% to 66.7% (while at the same time only increasing median false positive rate by 16% from 8% to 24%).

5.1.3 Initial False Alarms Results

IFA is the number of false positives a programmer must suffer through before they find a real security bug report. Table 9 shows our IFA results. There are three points to note from this table:

– FARSEC has no results in this table because FARSEC does not report results for this metric.
– For IFA, methods that only with/tune the data pre-processors perform worse than methods that optimize the learners (i.e., DE+Learners and SWIFT).
– In terms of absolute numbers, the IFA results are low for the Derby project. From Table 6, we can conjecture a reason for this – of the data set with a higher percentage of

security bug reports, the data sets have the more known target class, which is more likely to reduce the number of false positives encounter before the first correct prediction.
– At the other end of the spectrum, IFA is much larger for the Chromium project (median values for DE+learner or SWIFT of about 40 or 60). This result highlights the high cost of building highly secure software. When the target class is rare, even with our best-of-breed methods, some non-trivial amount of manual effort may be required.

5.1.4 Precision and F-Measure Results

For the sake of completeness, we also provide the results of precision and f-measure. Tables 10 and 11 present the corresponding precision and f-measure results from each technique besides FARSEC. We make the following remarks about these results.

– The decreasing trends are expected, as we select g-measure as our optimization target, which increases the recall and sacrifices the precision value. But, to some extent, these results also confirm the correctness of our choice. On the one hand, the improvement of recall with SWIFT is significant. On the other hand, for 4 out of 5 projects, the sacrifice of precision is moderate. For tasks such as bug report classification with the imbalanced data characteristic, as well in the context of security, in general, positive examples such as security bug reports are preferred not to be missed out. Hence, we would still recommend optimizing g-measure for future studies.
– There is little information gain in exploring both precision and f-measure since these results nearly echo each other (reason: f-measure is calculated as a combination of precision and recall).
– We admit the importance of precision, however, in some special domains such as security, there is little information gain in exploring precision results. As seen from our results, none of the techniques (including FARSEC) performs well under the precision metric. Hence, a low precision is not necessarily a reason to "discount" an optimizer. When the target class is rare, such low precision might actually be expected. For example, consider a query in the Google search engine, where it takes three pages before the user finds the target page. With 10 results per page, this means that the Google search engine is scoring a precision of $\frac{1}{30} \approx 3\%$. In this case, as precision is the fraction of retrieved pages that are relevant, such low precision is only a problem of time cost since the user wastes much time exploring irrelevant results before finding the target they care about. Our IFA results in Table 9, from the aspect of effort, also shows that, in the case of bug report classification, these low precision results *do not* lead to too much wasted time (evidence: the last row of Table 9 shows that users need to explore 15 to 22 false positives before finding a real security bug report – which is a small number when considering the size of total bug reports).

5.2 RQ2

> **RQ2.** When learning how to distinguish security bug reports, is it better to dual optimize the learners and the data pre-processors?

This research question explores the merits of dual optimization of learner plus pre-processor versus just optimizing one or the other. To answer this question, we count how

Table 10 **RQ1** results: precision. Higher values are better. Same as format as Table 7; i.e. best results are shown in boldface

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
Chromium	train	**31.0**	3.6	1.5	2.2	1.9
	farsecsq	**23.9**	3.3	1.4	1.8	2.7
	farsectwo	**31.0**	3.4	1.5	1.8	1.6
	farsec	**31.0**	3.6	1.2	2.6	2.8
	clni	**27.7**	3.8	1.6	2.8	1.7
	clnifarsecsq	**6.7**	2.7	1.5	1.8	2.8
	clnifarsectwo	**27.7**	3.8	1.4	3.1	2.2
	clnifarsec	**27.7**	2.4	1.5	2.3	1.9
	Median Prec	27.7	3.5	1.5	2.3	2.1
Wicket	train	2.8	0.0	2.5	**6.3**	2.2
	farsecsq	**2.1**	1.4	1.4	1.5	1.5
	farsectwo	**2.2**	1.4	1.2	1.3	1.3
	farsec	**4.8**	3.4	1.8	3.0	3.4
	clni	0.0	**8.3**	2.8	4.7	4.1
	clnifarsecsq	**1.6**	1.2	1.2	1.2	1.4
	clnifarsectwo	1.4	**1.5**	1.3	1.3	**1.5**
	clnifarsec	**5.5**	3.4	2.0	3.4	3.5
	Median Prec	2.2	1.5	1.6	2.3	1.9
Ambari	train	11.1	**40.0**	2.9	5.4	5.4
	farsecsq	4.1	**18.8**	2.0	3.4	4.1
	farsectwo	**21.1**	18.8	2.7	3.3	4.9
	farsec	4.0	**25.0**	3.0	7.9	4.8
	clni	7.1	**40.0**	4.7	6.5	5.3
	clnifarsecsq	9.5	**21.4**	4.3	7.9	2.7
	clnifarsectwo	8.3	**18.8**	4.5	10.7	4.3
	clnifarsec	**100.0**	21.4	7.3	13.6	5.1
	Median Prec	8.9	21.4	3.7	7.2	4.9
Camel	train	10.5	**30.0**	3.6	3.9	11.6
	farsecsq	5.2	5.6	6.7	8.2	**8.3**
	farsectwo	4.3	**7.7**	2.8	3.8	4.4
	farsec	**8.3**	4.8	2.6	3.6	5.5
	clni	4.8	**7.3**	3.0	4.0	7.0
	clnifarsecsq	4.3	**9.0**	7.8	6.2	7.1
	clnifarsectwo	**5.1**	4.0	3.7	3.8	**5.1**
	clnifarsec	**11.1**	5.2	4.0	4.1	5.3
	Median Prec	5.2	6.4	3.7	4.0	6.3
Derby	train	**34.0**	9.6	17.9	20.3	23.7
	farsecsq	14.4	11.5	8.5	10.6	**21.4**

Table 10 (continued)

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+ Pre-processors	Tune both (dual) SWIFT
	farsectwo	**26.0**	17.9	13.0	15.5	15.3
	farsec	35.6	**51.4**	19.3	21.6	30.0
	clni	**83.3**	52.9	24.0	18.2	19.4
	clnifarsecsq	14.4	**17.9**	8.6	8.6	12.7
	clnifarsectwo	**26.3**	11.0	12.1	15.3	10.5
	clnifarsec	34.0	**52.8**	16.0	31.9	23.9
	Median Prec	30.2	17.9	14.5	16.9	20.4
Overall median Prec		8.9	6.4	3.7	4.0	4.9

often each method achieves top-rank (and has boldface results) across all three metrics of the rows in Tables 7, 8 and 9.

Those count results are shown in Table 12. From this table, we can say, in terms of recall:

- SWIFT's dual optimization is clearly the best.
- Optimize just the data pre-processors comes a distant second.
- Optimize just the learners (with DE+Learners) is even worse.

Hence we say that, when distinguishing security bug reports, it is not enough to just tune the learners.

In terms of false positive rates, we see that:

- Optimize just the learner is a comparatively better method than other methods.
- Other treatments do not do well on the false alarm scale.

That said, optimize just the learner achieves a score of 14/40 – which is not even half the results. Hence, based on false positive rates, we cannot comment on what works best for improving this metric.

In terms of IFA (initial false alarms), we see that:

- Methods that do not optimize a learner do not perform well.
- There is is no clear winner for the best method since DE+Learners or SWIFT perform nearly the same as each other.

Based on the above observations, we could sum up the conclusions:

- Our experiment results show that dual optimization works well for recall.
- Also, not optimizing the learners performs badly for IFA.
- There is no clear pattern in Table 12 regarding false positive rates.

That said, the results of false positive rates seen in Table 8 are somewhat lower than the false positive rates seen in other software analytic papers (Menzies et al. 2006). Hence, on a more positive note, we can still recommend dual optimization since:

- It has many benefits (much higher recalls).

Table 11 **RQ1** results: f-measure

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+Pre-processors	Tune both (dual) SWIFT
Chromium	train	**20.8**	6.7	3.0	4.3	3.8
	farsecsq	**18.3**	6.2	2.7	3.5	5.2
	farsectwo	**20.8**	6.2	3.0	3.5	3.2
	farsec	**20.8**	6.6	2.4	5.0	5.4
	clni	**20.0**	6.8	3.2	5.5	3.3
	clnifarsecsq	**11.9**	5.3	2.9	3.6	5.3
	clnifarsectwo	**20.0**	7.0	2.8	6.0	4.3
	clnifarsec	**20.0**	4.6	2.9	4.5	3.8
	Median f-score	20.0	6.4	2.9	4.4	4.1
Wicket	train	4.8	0.0	4.8	**11.6**	4.2
	farsecsq	**4.0**	2.6	2.8	2.9	2.9
	farsectwo	**4.2**	2.8	2.3	2.5	2.6
	farsec	**8.3**	6.5	3.5	5.8	6.4
	clni	0.0	**11.1**	5.3	8.6	7.5
	clnifarsecsq	**3.0**	2.4	2.4	2.4	2.7
	clnifarsectwo	2.8	2.9	2.6	2.6	**3.0**
	clnifarsec	**9.8**	6.5	4.0	6.5	6.7
	Median f-score	4.1	2.8	3.2	4.4	3.6
Ambari	train	12.5	**33.3**	5.5	9.5	10.1
	farsecsq	7.4	**26.1**	3.8	6.4	7.8
	farsectwo	**30.8**	26.1	5.1	6.2	9.2
	farsec	6.3	**31.6**	5.6	13.3	8.9
	clni	9.5	**33.3**	8.5	11.3	9.9
	clnifarsecsq	16.3	**28.6**	7.9	13.3	5.2
	clnifarsectwo	12.9	**26.1**	8.1	17.1	8.1
	clnifarsec	25.0	**28.6**	12.5	20.7	9.5
	Median f-score	12.7	28.6	6.8	12.3	9.1
Camel	train	10.8	**21.4**	6.5	7.1	19.2
	farsecsq	7.9	9.7	11.4	14.3	**14.7**
	farsectwo	7.9	**12.8**	5.4	7.1	8.2
	farsec	**11.1**	7.5	4.7	6.4	10.0
	clni	7.5	10.2	5.4	7.2	**12.3**
	clnifarsecsq	6.8	**14.6**	12.2	10.2	12.6
	clnifarsectwo	7.0	7.4	7.0	7.2	**9.3**
	clnifarsec	**13.3**	7.9	7.0	7.4	9.7
	Median f-score	7.9	10.0	6.8	7.2	11.2
Derby	train	**36.0**	15.8	26.7	30.0	35.0
	farsecsq	22.8	19.1	14.7	18.4	**32.4**

Table 11 (continued)

Project	Filter	Prior state of the art (Peters et al. 2018) FARSEC	Optimize learners (only) DE+ Learners	Data pre-processing (no tuning) Pre-processors	Data pre-processing (tuned) DE+Pre-processors	Tune both (dual) SWIFT
	farsectwo	**33.6**	27.3	20.2	25.1	25.6
	farsec	36.8	**48.1**	28.6	31.4	40.9
	clni	37.0	**47.4**	33.8	27.9	30.1
	clnifarsecsq	22.8	**27.3**	15.4	15.2	21.3
	clnifarsectwo	**30.3**	18.5	19.8	24.4	18.1
	clnifarsec	36.0	**48.7**	25.3	40.4	35.2
	Median f-score	34.8	27.3	22.8	26.5	31.3
Overall median F-score		12.7	10.0	6.8	7.2	9.1

F-measure (or f-score) is defined as the harmonic mean of the model's precision and recall. Higher values are better. Same as format as Table 7; i.e. best results are shown in boldface

– With no excessive cost (not large increase in false alarms; IFA results are nearly as good as other methods).

Further to this comment of "no excessive cost", Table 13 shows the average runtime for each treatment. From the table, optimization on learners with the differential evolution algorithm consumes much more CPU time than others, while dual optimization as SWIFT shows slight advantages even better than optimizing data pre-processors. In addition, during our experiment, we also notice that, learners such as K Nearest Neighbors and Multilayer Perceptron can be slow to optimize, especially for large datasets such as the Chromium project which has about 20,000 data instances. However, since these learners are rarely selected as a "best" learner (see from Table 14), we would further recommend not using those learners for bug report classification task.

Table 12 How often is each treatment seen to be best in Tables 7, 8 and 9

Metric	Rank	Method	Win Times
Recall	1	SWIFT	31/40
	2	Pre-processors	14/40
	3	DE+Pre-processors	13/40
	4	DE+Learners	3/40
False Positive Rate	1	DE+Learners	14/40
	2	Pre-processors	1/40
	3	SWIFT	1/40
	4	DE+Pre-processors	0/40
IFA	1	DE+Learners	22/40
	2	SWIFT	18/40
	3	DE+Pre-processors	4/40
	4	Pre-processors	3/40

Table 13 Average runtime (in minutes) of optimizing all learner's hyperparameters, pre-processor's hyperparameters and running SWIFT

Project	DE3	DE10	Data Pre-processor Optimization	SWIFT
Chromium	455	876	20	12
Wicket	8	11	8	5
Ambari	8	11	8	5
Camel	8	11	8	5
Derby	8	11	8	5

Note that DE3 terminates after 3 generations and DE10 terminates after 10 generations

5.3 RQ3

RQ3. Can hyperparameter optimization further improve the performance of ranking security bug reports?

As the users of the bug reports, one of the major requirements is to distinguish as many actual security bug reports as possible. Our previous treatments are trying to seek a balance between recall and specificity, as stated in Section 4.4. The result of choosing *g-measure* as the optimization target is an increment of recall while at the cost of increasing false positive rate at the same time (see Tables 7 and 8). This usually could indicate that developers who use such tools would need to spend more time and effort to check those unexpected false positive predictions.

For many prominent applications such as web search engine, what is germane to users is how many good results are on the first page or the first two or three pages. Inspired by this, a ranking result of predicted bug reports would therefore be more helpful and reduce the required effort for developers. As we describe in Section 2.2, FARSEC employs a ranking method that sorts the predicted security bug reports. As a result, the actual security bug reports are closer to the top of the rank list.

We apply the same ranking technique as FARSEC, while the learners and/or pre-processors are optimized. The evaluation results based on the MAP metric are shown in Fig. 2. In the figure, the baseline (shown in blue color) is the method that does not apply any ranking technique (i.e., with the original chronological order). The orange line denotes

Table 14 The time of each learner that is selected as the "best" learner

Learner	FARSEC	DE+ Learners	Data Pre-processor	Data Pre-processor Optimization	SWIFT
Naive Bayes	6	21	23	16	17
Logistic Regression	16	5	5	4	3
Multilayer Perceptron	6	3	0	9	6
Random Forest	10	9	12	11	13
K Nearest Neighbors	2	2	0	0	1

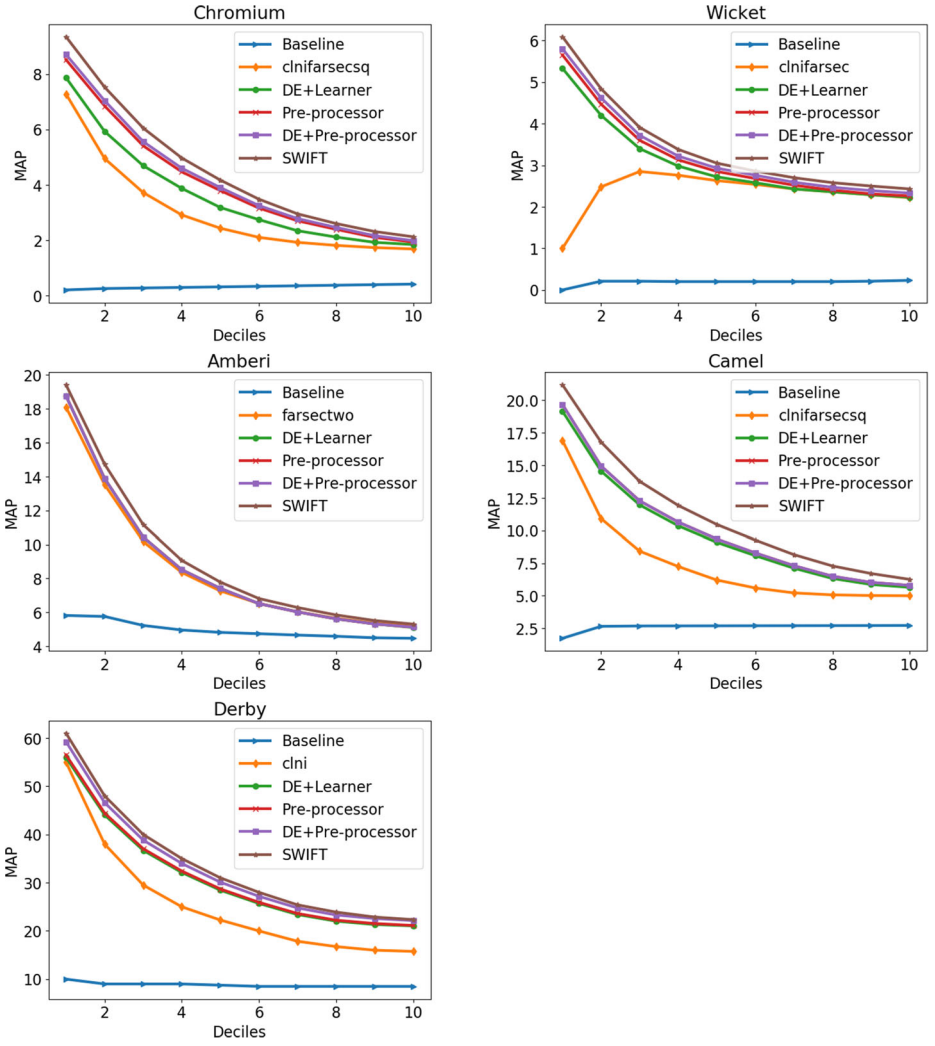

Fig. 2 Comparison of different treatments in ranking bug report prediction results. This plot shows its results using the deciles of (3) (from Section 4.4) and *higher* y-axis is *better*. Different treatments are denoted with lines of different colors. Specifically, the baseline (shown in *blue* color) is the method that does not apply any ranking technique (i.e., with the original chronological order). The *orange line* denotes the best ranking results from FARSEC among all filters

the best ranking results from FARSEC among all filters. The other treatments are denoted with lines of different colors.

The key observations from the figure are:

– The baseline method performs badly (the blue line) since this is with no ranking technique, whatsoever.
– In all data sets, the ranking generated using the prior state-of-the-art (the orange line for FARSEC filters) is below other treatments that try to rank the predicted bug reports.

– In a result that is consistent with the main message of this paper, in all data sets, the rankings generated by dual optimization (the brown SWIFT line) is above other methods.

The experiment results of ranking security bug reports, as well as results in previous research questions, could indicate that our proposed dual optimization of learners and pre-processors are promising. This approach could be recommended to better aid practitioners with similar domain tasks.

6 Discussion

SWIFT has demonstrated new results that improve the prior state-of-the-art. Speaking more broadly, what are the other lessons that could be taken from this work? We make the following comments.

Firstly, at the general application level, we have shown here it is possible to reason about rare event data (e.g., here the target security bug reports can be as rare as only taking up 1% of the total bug reports). Apart from the security case studied here, another lesson we would offer is that (sometimes) practitioners do not need (much) data to start data mining. This is an intriguing statement, since in this era of "big data", it is often assumed that scare of data would be a large obstacle. Here we offer a somewhat more optimistic comment: *effective models can be built even when data is scarce*.

Secondly, at the methodological level, we offer the following suggestion: *avoid using AI tools "off-the-shelf" without modifying them for the local domain*. SE practitioners need to develop specialized machine learning tools that are better suited to particular SE problems. Existing machine learning algorithms that we might call "general AI machine learning tools" maybe not "general" at all. Rather, they are tools whose default settings were chosen according to the data used *in the past* to commission those tools. Hyperparameter optimization tools should always be applied to adjust AI tools to the local data.

(Aside: One objection to the above point is that such optimization process can be unduly expensive. This objection is certainly true when we use traditional hyperparameter optimizers (e.g. genetic algorithms that evaluate thousands to millions of options (Holland 1992)). However, our empirical results from Table 13 shows that effective hyperparameter optimization can be accomplished in minutes. We note that, aside from data mining for security, previous researchers have achieved similar "fast optimization" results in several other SE domains (Agrawal et al. 2019).)

We are not the only researchers who make this second point. Other researchers in the software analytics literature also advocate tuning general AI tools to SE tasks. For example, Binkley et al. (2018) note that information retrieval tools for SE often equate word frequency with word importance, even though the number of occurrences of a variable name such as "tmp" is not necessarily indicative of its importance. They argue that the negative impacts of such differences manifest themselves when "off-the-shelf" information retrieval tools are applied in the software domain. Another example comes from sentiment analysis. Standard sentiment analysis tools are usually trained on non-SE data (e.g., the Wall Street Journal or Wikipedia). Novielli et al. (2018) recently developed their own sentiment analysis for the software engineering domain. After re-training those tools on an SE corpus, they found not only better performance at predicting sentiment, but also more agreement between different sentiment analysis tools.

Thirdly, it is natural to ask whether optimizing data pre-processors is more important than optimizing the learners (or vice versa). In reply, we say that there is no evident hints

from our empirical results show that one of them has obvious advantages over the other. In fact, recalling **RQ2**, we say that (at least in this domain) it is better to tune *both*.

Fourthly, another question we are asked is "in other domains, do our results say that some learners/pre-processors will perform better?". Our results do not support such conclusion. Table 14 shows that the "best" classifier is highly variable across our datasets. Hence, we cannot offer one general conclusion for all projects. However, what we do offer is a general method for finding the best local solution. Further, as shown by the runtime in Table 13, it may not be especially slow to apply our general method for finding the best local solution.

7 Threats to Validity

As to any empirical study, biases can affect the final results. Therefore, conclusions drawn from this work must be considered with threats to validity in mind.

Sampling Bias Sampling bias threatens any classification experiment. For example, the data sets used here come from FARSEC, i.e., one Chromium project and four Apache projects in different application domains. In addition, the bug reports from Apache projects are randomly selected with a BUG or IMPROVEMENT label for each project with extra labeling effort.

Learner Bias Research into automatic classifiers is a large and active field. While different machine learning algorithms have been developed to solve different classification problem tasks. Any data mining study, such as this paper, can only use a small subset of the known classification algorithms. For this work, we selected our learners such that we can compare our results to prior work. Accordingly, we used the same learners as Peters et al. in their FARSEC research.

Input Bias Our results come from the space of hyperparameter optimization explored in this paper. In theory, other ranges might lead to other results. That said, our goal here is not to offer the *best* optimization but to argue that *dual* optimization of data pre-processors and learners is preferable to optimize either, just by itself. For those purposes, we would argue that our current results suffice.

Evaluation Bias In our work, we choose some commonly used metrics as FARSEC for evaluation purpose and set *g-measure* as our optimization target. We do not use some other metrics because relevant information is not available to us or we think they are not suitable enough to this specific task (e.g., precision). In addition, we use equal weight in recall and specificity in the definition of g-measure, which is widely adopted in existing literature. We agree that it is important for these two elements to be re-weighted for different tasks, and this can be further explored as one of our future directions. Our implementation is flexible and we can adjust to proper metrics or balances with minor code modification.

8 Conclusion

Distinguishing security bug reports from other kinds of bug reports is a pressing problem that threatens not only the viability of software services, but also consumer confidence in those services. Prior results on how to distinguish security bug reports have had issues with

the scarcity of target data (specifically, such incidents occur rarely). In a recent TSE'18 paper, Peters et al. proposed some novel filtering algorithms to help improve security bug report classification. Results from FARSEC show that such filtering techniques can improve the performance.

But more than that, our experiments show that we can further do better than FARSEC using hyperparameter optimization of data mining learners and data pre-processors. Our results show that it is more advantageous to apply *dual* optimization of *both* the data-processor *and* the learner, which we will recommend in solving similar problems in future work.

Acknowledgments This work was partially funded via an NSF-CISE grant #1909516.

References

Agrawal A, Menzies T (2018) Is "Better Data" Better than "Better Data Miner"? (on the benefits of tuning SMOTE for defect prediction). In: Proceedings of the 40th international conference on software engineering, ACM, pp 1050–1061

Agrawal A, Fu W, Menzies T (2018) What is wrong with topic modeling? and how to fix it using search-based software engineering. Inf Softw Technol 98:74–88

Agrawal A, Fu W, Chen D, Shen X, Menzies T (2019) How to "DODGE" complex software analytics. IEEE Trans Softw Eng

Arcuri A, Briand L (2011) A practical guide for using statistical tests to assess randomized algorithms in software engineering. In: Proceedings of the 33rd international conference on software engineering ICSE '11. ACM, New York, pp 1–10. https://doi.org/10.1145/1985793.1985795

Bennin KE, Keung JW, Monden A (2019) On the relative value of data resampling approaches for software defect prediction. Empir Softw Eng 24(2):602–636

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(Feb):281–305

Bergstra JS, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Advances in neural information processing systems, pp 2546–2554

Biedenkapp A, Eggensperger K, Elsken T, Falkner S, Feurer M, Gargiani M, Hutter F, Klein A, Lindauer M, Loshchilov I et al (2018) Hyperparameter optimization. Artif Intell 1:35

Binkley D, Lawrie D, Morrell C (2018) The need for software specific natural language techniques. Empir Softw Eng 23(4):2398–2425

Black PE, Badger L, Guttman B, Fong E (2016) Dramatically reducing software vulnerabilities. Report to the White House Office of Science and Technology Policy, Information Technology Laboratory

Chan S, Treleaven P, Capra L (2013) Continuous hyperparameter optimization for large-scale recommender systems. In: 2013 IEEE international conference on big data, IEEE, pp 350–358

Chen L et al (2013) R2fix: automatically generating bug fixes from bug reports. Proceedings of the 2013 IEEE 6th ICST

Deb K, Mohan M, Mishra S (2005) Evaluating the ϵ-domination based multi-objective evolutionary algorithm for a quick computation of pareto-optimal solutions. Evol Comput 13(4):501–525

Deshmukh J, Podder S, Sengupta S, Dubash N et al (2017) Towards accurate duplicate bug retrieval using deep learning techniques. In: 2017 IEEE international conference on software maintenance and evolution (ICSME). IEEE, pp 115–124

Di Francescomarino C, Dumas M, Federici M, Ghidini C, Maggi FM, Rizzi W, Simonetto L (2018) Genetic algorithms for hyperparameter optimization in predictive business process monitoring. Inf Syst 74:67–83

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton

Feurer M, Springenberg JT, Hutter F (2015) Initializing bayesian hyperparameter optimization via meta-learning. In: Twenty-Ninth AAAI conference on artificial intelligence

Fu W, Menzies T (2017) Easy over hard: A case study on deep learning. In: Proceedings of the 2017 11th joint meeting on foundations of software engineering. ACM, pp 49–60

Fu W, Menzies T, Shen X (2016) Tuning for software analytics: is it really necessary? Inf Softw Technol 76:135–146

Gegick M, Rotella P, Xie T (2010) Identifying security bug reports via text mining: An industrial case study. In: 2010 7th IEEE working conference on mining software repositories (MSR). IEEE, pp 11–20
Goldberg DE (2006) Genetic algorithms. Pearson Education India
Goseva-Popstojanova K, Tyo J (2018) Identification of security related bug reports via text mining using supervised and unsupervised classification. In: 2018 IEEE international conference on software quality, reliability and security (QRS). IEEE, pp 344–355
Graham P (2004) Hackers & painters: big ideas from the computer age. O'Reilly Media, Inc
Han X, Yu T, Lo D (2018) Perflearner: learning from bug reports to understand and generate performance test frames. In: Proceedings of the 33rd ACM/IEEE international conference on automated software engineering. ACM, pp 17–28
Herodotou H, Lim H, Luo G, Borisov N, Dong L, Cetin FB, Babu S (2011) Starfish: a self-tuning system for big data analytics. In: Cidr, vol 11, pp 261–272
Hindle A, Alipour A, Stroulia E (2016) A contextual approach towards more accurate duplicate bug report detection and ranking. Empir Softw Eng 21(2):368–410
Holland JH (1992) Genetic algorithms. Sci Am 267(1):66–73
Huang Q, Xia X, Lo D (2017) Supervised vs unsupervised models: A holistic look at effort-aware just-in-time defect prediction. In: 2017 IEEE international conference on software maintenance and evolution (ICSME). IEEE, pp 159–170
Huang Q, Xia X, Lo D (2019) Revisiting supervised and unsupervised models for effort-aware just-in-time defect prediction. Empir Softw Eng 24(5):2823–2862
Jalali O, Menzies T, Feather M (2008) Optimizing requirements decisions with keys. In: Proceedings of the 4th international workshop on predictor models in software engineering. ACM, pp 79–86
Kampenes VB, Dybå T, Hannay JE, Sjøberg DIK (2007) A systematic review of effect size in software engineering experiments. Inf Softw Technol 49(11–12):1073–1086
Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. IEEE Trans Sys Man Cybern (4)580–585
Kim S, Zhang H, Wu R, Gong L (2011) Dealing with noise in defect prediction. In: 2011 33rd international conference on software engineering (ICSE). IEEE, pp 481–490
Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
Kochhar PS, Xia X, Lo D, Li S (2016) Practitioners' expectations on automated fault localization. In: Proceedings of the 25th international symposium on software testing and analysis. ACM, pp 165–176
Lamkanfi A, Demeyer S, Giger E, Goethals B (2010) Predicting the severity of a reported bug. In: 2010 7th IEEE working conference on mining software repositories (MSR). IEEE, pp 1–10
Lazar A, Ritchey S, Sharif B (2014) Improving the accuracy of duplicate bug report detection using textual similarity measures. In: Proceedings of the 11th working conference on mining software repositories. ACM, pp 308–311
Lessmann S, Baesens B, Mues C, Pietsch S (2008) Benchmarking classification models for software defect prediction: a proposed framework and novel findings. IEEE Trans Softw Eng 34(4):485–496
Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A (2017) Hyperband: a novel bandit-based approach to hyperparameter optimization. J Mach Learn Res 18(1):6765–6816
Menzies T, Shepperd M (2019) "Bad smells" in software analytics papers. Inf Softw Technol 112:35–47
Menzies T, Greenwald J, Frank A (2006) Data mining static code attributes to learn defect predictors. IEEE Trans Softw Eng 33(1):2–13
Menzies T, Dekhtyar A, Distefano J, Greenwald J (2007a) Problems with precision: a response to" comments on'data mining static code attributes to learn defect predictors'". IEEE Trans Softw Eng 33(9):637–640
Menzies T, Elrawas O, Hihn J, Feather M, Madachy R, Boehm B (2007b) The business case for automated software engineering. In: Proceedings of the Twenty-second IEEE/ACM international conference on automated software engineering ASE '07. ACM, New York, pp 303–312. https://doi.org/10.1145/1321631.1321676
Menzies T, Greenwald J, Frank A (2007c) Data mining static code attributes to learn defect predictors. IEEE Trans Softw Engineering (1) 2–13
Menzies T, Majumder S, Balaji N, Brey K, Fu W (2018) 500+ times faster than deep learning:(a case study exploring faster methods for text mining stackoverflow). In: 2018 IEEE/ACM 15Th international conference on mining software repositories (MSR). IEEE, pp 554–563
MITRE (2017) Common Vulnerabilities and Exposures (CVE). https://cve.mitre.org/about/terminology.html#vulnerability
Mittas N, Angelis L (2013) Ranking and clustering software cost estimation models through a multiple comparisons algorithm. IEEE Trans Softw Eng 39(4):537–551

Nair V, Yu Z, Menzies T, Siegmund N, Apel S (2018) Finding faster configurations using flash. IEEE Trans Softw Eng
Neuhaus S, Zimmermann T (2009) The beauty and the beast: vulnerabilities in red hat's packages. In: USENIX annual technical conference
Neuhaus S, Zimmermann T, Holler C, Zeller A (2007) Predicting vulnerable software components. In: Proceedings of the 14th ACM conference on computer and communications security. ACM, pp 529–540
Nguyen VH, Tran LMS (2010) Predicting vulnerable software components with dependency graphs. In: Proceedings of the 6th international workshop on security measurements and metrics. ACM, p 3
Novielli N, Girardi D, Lanubile F (2018) A benchmark study on sentiment analysis for software engineering research. In: 2018 IEEE/ACM 15Th international conference on mining software repositories (MSR). IEEE, pp 364–375
Ohira M, Kashiwa Y, Yamatani Y, Yoshiyuki H, Maeda Y, Limsettho N, Fujino K, Hata H, Ihara A, Matsumoto K (2015) A dataset of high impact bugs: manually-classified issue reports. In: 2015 IEEE/ACM 12th working conference on mining software repositories (MSR). IEEE, pp 518–521
Onan A, Korukoğlu S, Bulut H (2016) A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst Appl 62:1–16
Osman H, Ghafari M, Nierstrasz O (2017) Hyperparameter optimization to improve bug prediction accuracy. In: IEEE workshop on machine learning techniques for software quality evaluation (maLTeSQue). IEEE, pp 33–38
Panichella A, Dit B, Oliveto R, Di Penta M, Poshyvanyk D, De Lucia A (2013) How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms. In: International conference on software engineering
Parnin C, Orso A (2011) Are automated debugging techniques actually helping programmers? In: Proceedings of the 2011 international symposium on software testing and analysis. ACM, pp 199–209
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Peters F, Tun T, Yu Y, Nuseibeh B (2018) Text filtering and ranking for security bug report prediction. IEEE Trans Softw Eng:Early–Access
Scandariato R, Walden J, Hovsepyan A, Joosen W (2014) Predicting vulnerable software components via text mining. IEEE Trans Softw Eng 40(10):993–1006
Storn R, Price K (1997) Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359
Sun C, Lo D, Khoo SC, Jiang J (2011) Towards more accurate retrieval of duplicate bug reports. In: Proceedings of the 2011 26th IEEE/ACM international conference on automated software engineering. IEEE Computer Society, pp 253–262
Tantithamthavorn C, McIntosh S, Hassan AE, Matsumoto K (2016) Automated parameter optimization of classification techniques for defect prediction models. In: 2016 IEEE/ACM 38th international conference on software engineering (ICSE). IEEE, pp 321–332
Tantithamthavorn C, Hassan AE, Matsumoto K (2018) The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. IEEE Trans Softw Eng
The Equifax Data Breach (2019) https://epic.org/privacy/data-breach/equifax/
Thornton C, Hutter F, Hoos HH, Leyton-Brown K (2013) Auto-weka: combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 847–855
Tian Y, Lo D, Sun C (2012) Information retrieval based nearest neighbor classification for fine-grained bug severity prediction. In: 2012 19th working conference on reverse engineering. IEEE, pp 215–224
Tian Y, Lo D, Xia X, Sun C (2015) Automated prediction of bug report priority using multi-factor analysis. Empir Softw Eng 20(5):1354–1383
Van Aken D, Pavlo A, Gordon GJ, Zhang B (2017) Automatic database management system tuning through large-scale machine learning. In: Proceedings of the 2017 ACM international conference on management of data. ACM, pp 1009–1024
Vesterstrøm J, Thomsen R (2004) A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In: Congress on evolutionary computation. IEEE
Wang L, Zeng Y, Chen T (2015) Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. Expert Syst Appl 42(2):855–863
Wang Y, Xu W (2018) Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. Decis Support Syst 105:87–95
WannaCry Ransomware Attack (2017) https://en.wikipedia.org/wiki/WannaCry_ransomware_attack

Wijayasekara D, Manic M, McQueen M (2014) Vulnerability identification and classification via text mining bug databases. In: IECON 2014-40th annual conference of the IEEE industrial electronics society. IEEE, pp 3612–3618

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82

Xia X, Lo D, Qiu W, Wang X, Zhou B (2014) Automated configuration bug report prediction using text mining. In: 2014 IEEE 38Th annual computer software and applications conference (COMPSAC). IEEE, pp 107–116

Xia X, Lo D, Shihab E, Wang X (2016) Automated bug report field reassignment and refinement prediction. IEEE Trans Reliab 65(3):1094–1113

Xia Y, Liu C, Li Y, Liu N (2017) A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl 78:225–241

Yang X, Lo D, Huang Q, Xia X, Sun J (2016) Automated identification of high impact bug reports leveraging imbalanced learning strategies. In: 2016 IEEE 40Th annual computer software and applications conference (COMPSAC), vol 1. IEEE, pp 227–232

Yang XL, Lo D, Xia X, Huang Q, Sun JL (2017) High-impact bug report identification with imbalanced learning strategies. J Comput Sci Technol 32(1):181–198

Yildizdan G, Baykan ÖK (2020) A novel modified bat algorithm hybridizing by differential evolution algorithm. Expert Syst Appl 141:112949

Zaman S, Adams B, Hassan AE (2011) Security versus performance bugs: a case study on firefox. In: Proceedings of the 8th working conference on mining software repositories. ACM, pp 93–102

Zhang T, Yang G, Lee B, Chan AT (2015) Predicting severity of bug report by mining bug repository with concept profile. In: Proceedings of the 30th annual ACM symposium on applied computing. ACM, pp 1553–1558

Zhou Y, Sharma A (2017) Automated identification of security issues from commit messages and bug reports. In: Proceedings of the 2017 11th joint meeting on foundations of software engineering, pp 914–919

Zhou Y, Tong Y, Gu R, Gall H (2016) Combining text mining and data mining for bug report classification. J Softw Evol Process 28(3):150–176

Affiliations

Rui Shu[1] · Tianpei Xia[1] · Jianfeng Chen[1] · Laurie Williams[1] · Tim Menzies[1]

Rui Shu
rshu@ncsu.edu

Tianpei Xia
txia4@ncsu.edu

Jianfeng Chen
jchen37@ncsu.edu

Laurie Williams
lawilli3@ncsu.edu

[1] Department of Computer Science, North Carolina State University, Raleigh, NC 27695, USA