Developer-Driven Code Smell Prioritization

Fabiano Pecorelli,1 Fabio Palomba,1 Foutse Khomh,2 Andrea De Lucia1
1SeSa Lab - University of Salerno, Italy — 2École Polytechnique de Montréal, Canada
fpecorelli@unisa.it, fpalomba@unisa.it, foutse.khomh@polymtl.ca, adelucia@unisa.it

ABSTRACT
Code smells are symptoms of poor implementation choices applied during software evolution. While previous research has devoted effort in the definition of automated solutions to detect them, still little is known on how to support developers when prioritizing them. Some works attempted to deliver solutions that can rank smell instances based on their severity, computed on the basis of software metrics. However, this may not be enough since it has been shown that the recommendations provided by current approaches do not take the developer’s perception of design issues into account. In this paper, we perform a first step toward the concept of developer-driven code smell prioritization and propose an approach based on machine learning able to rank code smells according to the perceived criticality that developers assign to them. We evaluate our technique in an empirical study to investigate its accuracy and the features that are more relevant for classifying the developer’s perception. Finally, we compare our approach with a state-of-the-art technique. Key findings show that the our solution has an F-Measure up to 85% and outperforms the baseline approach.

KEYWORDS
Code smells; Machine Learning for Software Engineering; Empirical Software Engineering.

ACM Reference Format:
Fabiano Pecorelli,1 Fabio Palomba,1 Foutse Khomh,2 Andrea De Lucia1.

1 INTRODUCTION
As a software system evolves, continuous modifications are required to adapt it to new requirements and/or changing environments or even fix defects that can preclude its correct functioning [69]. These activities are usually performed between strict deadlines and constraints to meet [9]. As a side effect, developers risk to introduce technical debts [9, 78], i.e., sub-optimal implementation decisions that provide short-term benefits but cause a decrease of software quality. One of the main indicators of technical debt is represented by the presence of code smells [26], which are poor design/implementation choices applied by developers during maintenance and evolution, e.g., the introduction of complex and/or long classes, excessive coupling between objects, and so on.

Code smells have been often associated to a decrease of program comprehensibility [1], maintainability [38, 59], testability [28] as well as an increase of maintenance effort and costs [80]. These findings have motivated researchers to propose automated mechanisms to support developers in both the identification [4, 18, 21] and removal [51] of code smells, obtaining promising results.

Despite this effort, the adoption of code smell detectors in practice is still limited [57]. Among the others, two notable reasons precluding the applicability of automated detectors in practice may be: (1) the large amount of code smell instances detected by these tools combined to the little knowledge on how to prioritize their refactoring [25, 49]; (2) the empirical evidence that most of the available detectors identify code smells that developers do not perceive or do not consider critical [23, 65].

While some researchers provided initial attempts toward the prioritization of code smells using measures of severity derived from software metrics [3, 25, 49, 93], the available solutions either rely on predefined heuristics that have not been empirically assessed or do not address the problem of providing developers with recommendations aligning with their perception of design issues, thus possibly being still ineffective in practice.

In this paper, we build on this line of research and propose the first step toward the concept of developer-driven code smell prioritization as an alternative and more pragmatic solution to the problem: Rather than ranking code smell instances based on their severity computed using software metrics, we propose to prioritize them according to the criticality perceived by developers.

In particular, we first perform surveys to collect a dataset composed of developers’ perception of the severity of 1,332 code smell instances—pertaining to four different types of design problems—for which original developers rated their actual criticality and then propose a novel supervised approach that learns from such labeled data to rank unseen code smell instances. Afterwards, we conduct an empirical study to (1) verify the performance of our prioritization approach, (2) understand what are the features that contribute most to model the developer’s perceived criticality of code smells, and (3) compare our approach with the state-of-the-art baseline proposed by Arcelli Fontana and Zanoni [25]. The key differences between our approach and the baseline considered for the comparison are (i) the usage of different kinds of predictors (e.g., process metrics) rather than considering only structural ones, and (ii) the definition of a dependent variable based on the developers’ perception.

The main result of the study highlights that the devised approach can classify the developer’s perceived criticality of code smells with an F-Measure ranging between 72% and 85%. Moreover, we discovered that, depending on the code smell type, specific features are more relevant to classify its criticality. Finally, our approach...
performs better than the experimented baseline when classifying all considered code smells.

To sum up, this paper provides the following contributions:

1. A new dataset reporting the criticality perceived by original developers with respect to 1,332 code smell instances of four different types, which can be further used by the community to build upon our research;
2. The first machine learning-based approach to prioritize code smells according to the real developers’ perceived criticality;
3. An empirical study that showcases the performance of our approach, the importance of the employed features, and the reasons why our technique goes beyond the state-of-the-art;
4. An online appendix with the datasets used in the study, that can be exploited to replicate and extend our work.

Structure of the paper. In Section 2 we summarized the related literature. Section 3 describes the methods employed to construct the dataset, while Section 4 reports on the definition of our approach and its empirical evaluation. In Section 5 we discuss the results of the study, while Section 6 overviews the threats to the validity of the study and how we mitigated them. Finally, Section 7 concludes the paper and outlines our future research agenda.

2 RELATED WORK

The research community actively studied code smells in the past [4, 18, 21]. On the one hand, researchers focused on understanding the characteristics of code smells and, in particular, their origin [47, 60, 91], evolution [56], diffuseness [59], relevance for developers [61, 65, 85, 94, 95], impact on software comprehensibility [1, 66] and maintainability [6, 38, 80]. On the other hand, a number of automatic code smell detection approaches and tools have been developed and validated [20, 24, 37, 39, 52, 62, 64, 68, 71, 89].

When it comes to code smell prioritization, however, the research contribution so far is notably less prominent and much more focused on the idea of ranking refactoring recommendations. For instance, Tsantalis and Chatzigeorgiou [90] proposed to order the refactoring suggestions given by JDeodorant using historical information and, particularly, modeling the probability that a certain refactorable class will change in the near future. Along these lines, Girba et al. [27] employed the number of changes as metric to decide on whether a class should be refactored in the current release of a software system. With respect to these papers, there are two main aspects that make our paper different. First, there may be refactoring opportunities that do not target classes affected by code smells, i.e., many refactoring actions are driven by other considerations, e.g., the introduction of design patterns [6, 79]. As such, the goal of our paper is diametrically different, since we aimed at ranking code smells based on their perceived criticality. In the second place, the papers above are based on the underlying concept that the recent history of class is the main driver for its refactorability, while we showed that multiple metrics may possibly contribute to it.

Other researchers took a closer look to the problem of prioritizing code smells. Marinescu [49] defined a method to rank smell instances based on their severity that can be applied on top of heuristic-based code smell detectors: given a set of metrics that characterize a class and given the corresponding thresholds that discriminate high/low values for these metrics, the method computes the average distance between the actual code metric values and the fixed thresholds. That is, smell instances having higher average distance from the thresholds are ranked at the top. Later on, Arccoverde et al. [3] proposed heuristics to rank code anomalies based on their impact on the overall architecture of the system, computed on the basis of number of changes and faults. These two works are essentially based on predefined heuristics, hence implementing a different approach with respect to the one proposed in our paper. Also, these approaches have not been tested in practice and, for this reason, little is known about their actual capabilities. Nevertheless, in our study we considered these papers when defining the set of independent variables of our model: we considered both number of changes and faults coming from Arccoverde et al. [3] as well as the severity computed as suggested by Marinescu [49].

The two closest works are by Vidal et al. [93] and Arcelli Fontana and Zanoni [25]. The first approach takes into account three main factors to prioritize code smells, i.e., stability, relevance, and modifiability scenarios; however, it is semi-automated and requires the input of experts to be actually used. On the contrary, with this paper we aimed at providing a fully automated solution. As for Arcelli Fontana and Zanoni [25], the authors defined a machine learner based on a large variety of size, cohesion, coupling, and complexity metrics, to predict the severity of the presence of a certain code smell in a class/method (from low to high). In their empirical study, the approach reached up to 95% in terms of F-Measure. Our work built on top of the work by Arcelli Fontana and Zanoni [25] and proposed a further step ahead: first, we aimed at classifying the perceived criticality of code smells in order to define methods that are closer to what developers consider real design issues; second, we do that by exploiting metrics of different nature and able to characterize the perceived quality of classes under different perspectives.

3 DATASET CONSTRUCTION

To perform our empirical study, we needed to collect a dataset reporting the perceived criticality of a set of code smells large enough to train a machine learning model. To this aim, we first defined the objects of the study, namely (1) a set of software projects and (2) the code smell types we were interested in with their corresponding detectors; Then, we inquired the subjects of our study, namely the original developers of the considered projects, in order to collect their perceived criticality of the code smell instances detected on their codebase. The next subsections describe the various steps we followed to build our dataset.

3.1 Selecting projects

The context of the study consisted of nine open-source projects belonging to two major ecosystems such as Apache and Eclipse. Basic information and statistics about the selected projects are summarized in Table 1. Specifically, for each considered project, we report (i) the total number of commits available in its change history, (ii) the total number of contributors, and (iii) the size as the

1https://www.apache.org
2https://www.eclipse.org/org/
we focused on four class-level types of code smells, namely: written in Java. This smell affects classes that do not properly change history at least account those having a number of classes higher than available at the time of the analysis on codebase size, (b) domain, (c) longevity, (d) activity, and (e) population. As such, starting from the set of 2,576 open-source systems written in JAVA and belonging to the two considered ecosystems available at the time of the analysis on GITHUB, we only took into account those having a number of classes higher than 500, with a change history at least 5 years long, having at least 1,000 commits, and with a number of contributors higher than 20. This filter gave us a total of 682 systems: of these, we randomly selected 9 of them.

3.2 Selecting code smells

We focused on four class-level types of code smells, namely:

**Blob (or God Class).** This code smell type affects classes that do not follow the single-responsibility principle [88], i.e., they implement more than one responsibility, thus being poorly cohesive and hard to understand/maintain [26]. Previous studies demonstrated that classes affected by this smell are connected to higher change- and defect-proneness [38, 59] as well as maintenance and evolution costs [1, 80, 81]. According to recent findings [61, 85], this smell is among the most critical ones for practitioners.

**Complex Class.** Instances of this smell affect classes that have high cyclomatic complexity [10] and that, therefore, may primarily make the testing of those classes harder [28, 50], but also lower the ability of developers to evolve them [59]. Existing empirical evidence on the developer’s perception about this smell indicated that practitioners are generally able to recognize this smell and assess its criticality [61, 95].

**Spaghetti Code.** This smell affects classes that do not properly use Object-Oriented programming principles (e.g., inheritance), declare at least one long method with no parameters, and use instance variables [26] — it refers to classes that follow a functional programming style. Similarly to the other considered code smells, also Spaghetti Code has been widely investigated in the past by researchers, who discovered that it hinders the ability of developers to comprehend source code [1], increases maintenance effort [38, 80], and can be accurately assessed by developers with respect to its criticality [61, 85].

### Shotgun Surgery

When a change to a class (e.g., to one of its fields/methods) triggers many little changes to other classes of the system, such class is affected by this smell [26]. Instances of this smell are associated with a higher defect-proneness of the involved class [17]. Previous studies on the perception of this smell revealed that practitioners perceive its presence depending on the intensity of the flaw [61], i.e., depending on the number of changes triggered on other classes of the project.

Three specific factors drove the selection of these four code smell types. First, they have been shown to be highly diffused in real software systems [59], thus allowing us to target code smells that are relevant in practice. Second, they are reported to negatively impact maintainability, comprehensibility, and/or testability of software systems [28, 38, 59]; as such, we could investigate design flaws that practitioners may be more able to analyze and assess. Finally, previous findings [61, 85, 95] showed not only that they are actual problems from the developer’s perspective, but also that their criticality can be accurately assessed by practitioners, thus mitigating potential problems due to the presence of the so-called conceptual false positives [23], i.e., code smell instances detected as such by automated tools but not representing issues for developers.

### 3.3 Selecting code smell detectors

Once we had selected the specific code smells object of our investigation, we then proceeded with the choice of automated code smell detectors that could identify them. Among all the available solutions proposed so far by researchers [21], we opted for DECOR [52] and Hist [62]. The first was selected to identify instances of Blob, Complex Class, and Spaghetti Code, while the latter for the detection of Shotgun Surgery.

More specifically, DECOR is an automated solution which adopts a set of “rule cards”, namely rules able to describe the intrinsic characteristics that a class must have to be affected by a certain code smell type. In the case of Blob, the approach identifies it when a class has a Lack of Cohesion of Method (LCOM5) [34] higher than \( \alpha \), a total number of methods and attributes higher than \( \beta \), it is associated to many data classes (i.e., classes having just get and set methods), and has a name having a suffix in the set \{Process, Control, Command, Manage, Drive, System\}, where \( \alpha \) and \( \beta \) are relative threshold values. When detecting Complex Class instances, DECOR computes the Weighted Methods per Class metric (WMC), i.e., the sum of the cyclomatic complexity of all methods of the class [50], and marks a class as smelly if the WMC is higher than a defined threshold. Finally, Spaghetti Code instances are represented by classes presenting (i) at least one method without parameters and having a number of lines of code higher than a defined threshold, (ii) no inheritance, as indicated by the Depth of Inheritance Tree metric (DIT) [15] which must be equal to 1, and (iii) a name suggesting procedural programming, thus having as prefix/suffix a word in the set \{Make, Create, Exec\}. There are two key reasons leading us to rely on DECOR for the detection of these three smells. In the first place, this detector has been employed in several previous studies on code smells [36, 39, 46, 65, 66], showing good results

<table>
<thead>
<tr>
<th>Project</th>
<th>#Commits</th>
<th>#Devs</th>
<th>#Classes</th>
<th>KLOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apache Mahout</td>
<td>3,054</td>
<td>55</td>
<td>813</td>
<td>204</td>
</tr>
<tr>
<td>Apache Cassandra</td>
<td>2,026</td>
<td>128</td>
<td>586</td>
<td>111</td>
</tr>
<tr>
<td>Apache Lucene</td>
<td>3,784</td>
<td>62</td>
<td>5,506</td>
<td>142</td>
</tr>
<tr>
<td>Apache Cayenne</td>
<td>3,472</td>
<td>21</td>
<td>2,854</td>
<td>542</td>
</tr>
<tr>
<td>Apache Pig</td>
<td>2,432</td>
<td>24</td>
<td>826</td>
<td>372</td>
</tr>
<tr>
<td>Apache Jackrabbit</td>
<td>2,924</td>
<td>22</td>
<td>872</td>
<td>527</td>
</tr>
<tr>
<td>Apache Jena</td>
<td>1,489</td>
<td>38</td>
<td>663</td>
<td>231</td>
</tr>
<tr>
<td>Eclipse CDT</td>
<td>5,961</td>
<td>31</td>
<td>1,415</td>
<td>249</td>
</tr>
<tr>
<td>Eclipse CFX</td>
<td>2,276</td>
<td>21</td>
<td>655</td>
<td>106</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td><strong>32,889</strong></td>
<td><strong>436</strong></td>
<td><strong>5,506</strong></td>
<td><strong>542</strong></td>
</tr>
</tbody>
</table>
when considering both precision and recall. At the same time, it implements a lightweight mechanism with respect to other existing approaches (e.g., textual-based techniques relying on information retrieval [64, 67]): the scalability of DECOR allows us to perform an efficient detection on the large systems considered in the study.

Turning our attention to the detection of Shotgun Surgery, the discussion is different. Approaches based on source code analysis are poorly effective for the detection of this smell [62]; for instance, the approach proposed by Rao and Reddy [75]—which computes coupling metrics to build a change probability matrix that is then filtered to detect the smell—was not able to identify any Shotgun Surgery instance when applied in large-scale studies [62]. For this reason, we relied on Hist [62], a historical-based technique that (1) computes association rules [2] to identify methods of different classes that often change together and (2) identifies instances of the smell if a class contains at least one method that frequently changes with methods present in more than 3 different classes. Such a historical-based approach has shown an F-Measure close to 92% [62], thus representing a suitable solution for conducting our study.

On a technical note, we relied on the original implementations of DECOR and Hist, hence avoiding potential threats to construct validity due to re-implementations.

### 3.4 Collecting the criticality of code smells

The last step required to build our dataset was the actual detection of code smells in the considered systems and the subsequent inquiry on their criticality. We followed a similar strategy as Silva et al. [79]: in short, in a time period of 6 months, from January 1st to June 30th, 2018, we monitored the activities performed on the selected repositories and, as soon as a developer committed changes to classes affected by any of the considered smells, we sent an e-mail to that developer to ask (i) whether s/he actually perceived/recognized the presence of a code smell and (ii) if so, rate its criticality using a Likert scale from 1 (very low) to 5 (very high) [43].

In particular, we built an automated mechanism that fetched—using the git-fetch command—commits from the repositories to a local copy on a daily basis. This gave us the possibility to generate the list of classes modified during the workday. At this point, we performed the actual smell detection. In the case of DECOR, we parsed each modified class and run the detection rules described in Section 3.3 to identify instances of Blob, Complex Class, and Spaghetti Code. As for Hist, it requires information about the change history of the involved classes: for this reason, before running the Shotgun Surgery detection algorithm, we mined the log file of the projects to retrieve the set of changes applied on the classes modified during the workday. Through the procedure described so far, we obtained a list of smells classes, and, for each of them, we stored the e-mail address of the developer who committed changes on it. Afterwards, we manually double-checked the smelly classes given by the automated tools with the aim of discarding possible false positives, thus avoiding asking developers useless information. Overall, the code smell detection phase resulted in a total of 2,675 candidate code smells. Of these, we discarded 455 (=17%) false positives.

Finally, we sent e-mails to the original developers. In the text, we first presented ourselves and then explained that our analysis tool suggested that the developer likely worked on a class affected by a design issue—without revealing the exact code smell to avoid confirmation bias [55]. Then we asked three specific questions:

1. **Were you aware of the presence of a design flaw?**
2. **If yes to question (1), may you please briefly describe the type of design flaw affecting the class?**
3. **If yes to question (1), may you please rate the criticality of the design flaw from 1 (very low) to 5 (very high)?**

We asked the first question to make sure that the contacted developers perceived classes as affected by code smells. If not, they could not obviously provide meaningful information on their criticality, and the answers were discarded. Otherwise, we further asked to explain the design problem perceived, so that we could understand if developers were actually aware of the specific smell. When receiving the answers, we checked if the explanation given by developers was in line with the definition of the smell: for instance, one developer was contacted to rate the criticality of a Blob class and explained that “[the class] is a well-known problem, it is huge in size and has high coupling”, thus indicating that s/he correctly recognized the smell we were proposing to him/her. In these cases, we considered the answer to the third question valid, otherwise we discarded it. Note that if a code smell was detected in the same class more than once, we did not send any e-mail to not bother the developers multiple times for the same class.

As an outcome, we sent a total of 1,733 e-mails to 372 distinct developers, i.e., an average of 0.77 e-mails per month per developer, while 487 code smells affected the same classes multiple times and, therefore, we avoided sending e-mails for them. Moreover, we could not assess the criticality of 310 code smells because the 139 developers responsible for them did not reply to our e-mails. Also, we had to discard 91 answers received since the contacted developers did not perceive the presence of code smells, i.e., they answered ‘no’ to question (1).

Hence, we finally gathered 1,332 valid answers coming from 233 developers: the high response rate (62%) is in line with previous works that implemented a similar recruitment strategy [66, 79] and indicates that contacting developers immediately after their activities with short surveys not only increases the chances of receiving accurate answers [79], but also helps increasing their overall responsiveness. As a final note, the 1,332 code smell instances evaluated were almost equally distributed among the four considered types of design flaw: indeed, we had answers for 341 Blob, 349 Complex Class, 313 Spaghetti Code, and 329 Shotgun Surgery instances. Also, the criticality values assigned by developers to each smell type were almost uniformly distributed over the possible ratings (1 to 5)—more details are available in our online appendix [72].

### 4 A NOVEL CODE SMELLS PRIORITIZATION APPROACH AND ITS EVALUATION

The goal of our study is to define and assess the feasibility of using a machine learning-based solution to prioritize code smells according to their perceived criticality, with the purpose of providing developers with recommendations that are more aligned to the way they actually refactor source code. The perspective is of both practitioners and researchers: the former are interested in adopting more practical solutions to prioritize refactoring activities, while
the latter are interested in assessing how well machine learning can be employed to model developer’s criticality of code smells.

4.1 Research Questions
The empirical study revolves around three main research questions (RQs). We started with the definition of a machine learning-based approach to model the developer’s perceived criticality of code smells. Starting from the dataset built following the strategy reported in Section 3, we defined dependent and independent variables of the model as well as the appropriate machine learning algorithms to deal with the problem. These steps led to the definition of our first research question:

**RQ1. Can we predict developer’s perception of the criticality of a code smell?**

Besides assessing the model as a whole, we then took a closer look at the contributions given by the independent variables exploited, namely what are the features that help the most when classifying the perceived criticality of code smells. This step allowed us to verify some of the conjectures made when defining the set of independent variables to be used. Hence, we asked:

**RQ2. What are the features of the proposed approach that contribute most to its predictive performance?**

As a final step of our investigation, we considered the literature in the field to identify existing code smell prioritization approaches that can be used as baselines, thus allowing us to assess how useful our technique can be when compared to existing approaches. This led to our final research question:

**RQ3. How does our approach perform when compared with existing code smell prioritization techniques?**

In the next sections, we describe the methodological details of the evaluation of the proposed approach.

4.2 Defining and assessing the performance of the prioritization approach
To address RQ1, we defined a novel code smell prioritization approach that aims at classifying smell instances based on the developer’s perceived criticality using machine learning algorithms. This implied the definition of an appropriate set of independent variables able to predict the dependent variable, i.e., the developer’s perception, as well as the proper algorithms and their configuration.

**Dependent Variable.** As a first step, we defined the developer’s perceived criticality of code smells as a variable that the model has to estimate. The dataset described in Section 3 reports, for each code smell instance, a value ranging from 1 to 5 describing the perceived criticality of that instance. Thus, we mapped the problem as a classification one [16], namely we took into account the case in which the learner has to classify the criticality of code smells in multiple categorical classes [16]. In this case, we converted the integers of our dataset in nominal values in the set \{low, medium, high\}: if a code smell instance was associated to 1 or 2 in the original dataset, then we considered its perceived criticality as low; if it was equal to 3, we converted its value in medium; otherwise, we considered its criticality as high. With this mapping, we merged the values assigned by developers in order to build three main classes. This was a conscious decision given by experimental tests: indeed, when experimenting with a 5-point classification problem, we observed that several misclassifications were due to the approach not able to correctly distinguish (i) very-low from low and (ii) high from very-high. Thus, we opted for a 3-point classification.

**Independent Variables.** To predict the perceived criticality of code smells, we considered a set of features able to capture the characteristics of classes under different angles. Table 2 summarizes the families of metrics considered, the rationale behind their use, and the specific indicators measured. Previous research has not only shown that product metrics can indicate actual design problems in source code [13, 53], but also lead developers to recognize the presence of sub-optimal implementation solutions that would deserve some refactoring [61, 84, 85]. For these reasons, we considered four types of product metrics, such as (i) size, (ii) cohesion, (iii) coupling, and (iv) complexity metrics. For each of these types, we selected indicators having different nature (e.g., structural aspects of source code rather than textual components) and able to capture in different ways the considered phenomena (e.g., we computed source code complexity using both the McCabe metric and readability, which targets a more cognitive dimension of complexity).

While product metrics can provide indications about the structure of source code, we complemented them with orthogonal metrics that capture the way the code has been modified as well as who was responsible for that, i.e., process and developer-related metrics. Indeed, the developer’s perception of criticality may be not always due to the complex structure of source code, but rather to the problems it causes during the evolution process [65, 74]; similarly, the criticality of code smells may be perceived differently depending on whether the maintainer is an expert of the class or not [8, 14]. Hence, we selected a number of metrics related to those aspects: for instance, we computed the average number of co-changing classes for the smell class (AVG_CS) to assess whether smells that often change with several classes are perceived as more critical by developers or the number of previous bug fixing involving the smelly class to observe if classes that are more fault-prone are actually those perceived more critical. We also computed measures of experience and ownership of developers working on the smelly class to test whether these factors influence the developer’s ability to work on it. Finally, we took into account some specific metrics related to code smells: the idea here is that a number of aspects connected to the smell itself may be relevant for developers when assessing its criticality. In particular, the continuous presence of smell over the history of the project (i.e., Pers.) may influence the ability of developers to recognize its harmfulness better. Much in the same way, the presence of refactoring opportunities (Ref) or even the number of previous refactoring actions done on the smelly class (NR) may affect the perception of developers. Finally, we also considered the code smell intensity, which is a measurable amount of intensity of a certain code smell instance [49]: we
Table 2: Software metrics used as independent variables split by categories - The motivations for their use are also reported.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines of Code</td>
<td>LOC</td>
<td>Amount of lines of code of a class excluding white spaces and comments.</td>
</tr>
<tr>
<td>Lack of Cohesion of Methods [15]</td>
<td>LOCMS</td>
<td>Number of method pairs in a class having no common attribute references.</td>
</tr>
<tr>
<td>Coupling between Object Classes [15]</td>
<td>CBO</td>
<td>Number of classes in the system that call methods or access attributes of a class.</td>
</tr>
<tr>
<td>Message Passing Coupling [15]</td>
<td>MPC</td>
<td>Number of method calls made by a class to external classes of the system.</td>
</tr>
<tr>
<td>Response for a Class [15]</td>
<td>RFC</td>
<td>Sum of the methods of the class, i.e., number of methods that can potentially be executed in response to a message received by an object of a class.</td>
</tr>
</tbody>
</table>

Process metrics. The amount of activities made on smelly classes may affect the developer’s perceived criticality [42, 45, 74].

<table>
<thead>
<tr>
<th>Metric</th>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Commit Size</td>
<td>AVG_CS</td>
<td>Average number of classes that co-changed in commits involving a class.</td>
</tr>
<tr>
<td>Number of Changes</td>
<td>NC</td>
<td>Number of commits in the change history of the system involving a class.</td>
</tr>
<tr>
<td>Number of Bug Fixes</td>
<td>NF</td>
<td>Number of bug fixing activities performed on a class in the change history of the system.</td>
</tr>
<tr>
<td>Number of Committers</td>
<td>NCOM</td>
<td>Number of distinct developers who performed commits on a class in the change history of the system.</td>
</tr>
</tbody>
</table>

Developer-related metrics. Experience and workload of developers may affect the perceived criticality of code smells [8, 12, 14, 91].

<table>
<thead>
<tr>
<th>Metric</th>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developer’s Experience [8]</td>
<td>EXP</td>
<td>Average number of commits of the committers of a class.</td>
</tr>
<tr>
<td>Developer’s Scattering Changes [19]</td>
<td>DSC</td>
<td>Average number of distinct subsystems in which the committers of a class made changes.</td>
</tr>
<tr>
<td>Development Change Entropy [33]</td>
<td>CE</td>
<td>Shannon’s entropy [77] computed on the number of changes of a class in the change history of the system.</td>
</tr>
<tr>
<td>Code Ownership [8]</td>
<td>OWN</td>
<td>Number of commits of the major contributor of a class over the total number of commits for that class.</td>
</tr>
</tbody>
</table>

Code smell-related metrics. Persistence of code smells and availability of refactoring opportunities may affect the developer’s perception [38, 68].

<table>
<thead>
<tr>
<th>Metric</th>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>Pers.</td>
<td>Number of subsequent major/minor releases in which a certain smell affects a class.</td>
</tr>
<tr>
<td>Intensity [49]</td>
<td>Int.</td>
<td>Average distance between the actual metric values used for the detection of code smells and the corresponding thresholds considered by the detectors to distinguish smelly and non-smelly elements.</td>
</tr>
<tr>
<td>Refactorable</td>
<td>Ref.</td>
<td>Existence of refactoring opportunities for a class, as detected by automated tools.</td>
</tr>
<tr>
<td>Number of Refactoring Actions</td>
<td>NR</td>
<td>Number of previous refactoring operations made by developers on a class.</td>
</tr>
</tbody>
</table>

Note that, before running the algorithms, we first applied a forward selection of independent variables using the CORRELATION-BASED FEATURE SELECTION (CFS) approach [44], which uses correlation measures and a heuristic search strategy to identify a subset of actually relevant features for a model, thus mitigating possible problems due to multi-collinearity of features [54]. Then, we configured classifiers’ hyper-parameters by exploiting the GRID SEARCH [7] algorithm, that runs an exhaustive search of the hyper-parameter space: in so doing, the algorithm applies a 10-fold cross-validation [40] in order to evaluate multiple times the effect of parameters on the performance of the model.

Training/Testing the Model. We built different models for each code smell considered in the study, so the training data is represented by the set of observations available for a certain smell in the collected dataset—the distribution of the criticality values assigned by developers to the code smell instances evaluated is available in our appendix [72]: the distribution is almost uniform for all the criticality values. This aspect affected our decision to not apply any balancing algorithm. On the one hand, there are no classes requiring to be balanced with respect to the others. On the other hand, previous findings [70, 71] have shown that balancing code smell-related datasets can even damage the performance of the resulting models.

To train the model, we employed a 10-fold cross-validation strategy [40]: it randomly partitions the available dataset into 10 folds of equal size, applying a stratified sampling—meaning that each fold has the same proportion of the various criticality classes. A single fold is then used as test set, while the remaining ones are employed for training the model. The process is repeated ten times so that each fold will be the test set exactly once.

Performance Assessment. We evaluated the performance of the experimented model by analyzing confusion matrices, obtained included this metric to understand whether there is a match between an "objective" measurement of code smell severity and its real perceived criticality.

From a technical perspective, we employed the tool by Spinellis [83] to compute product metrics, the one made available by Buse and Weimer [11] for the computation of the readability index, and PyDriller [82] to compute process and developer-related metrics. As for the smell-related indicators, we developed our own tool to compute Pers. and NR. Starting from the release $R_i$ of the projects taken into account, the former metric counts in how many consecutive previous major and minor releases—identified using the corresponding Git tags—the considered smell was present, according to the employed detectors. The latter metric, instead, was computed by mining the messages of commits involving the smelly classes and looking for the presence of keywords recommended in [92], e.g., ‘refactor’ or ‘restructure’. The intensity of code smells has been assessed using the tool by Marinescu [49], which computes the average distance between the actual code metric values of the smell instance and the thresholds fixed by the detection rules. Finally, the Ref. metric was computed by (i) running JDEODORANT [22], an existing refactoring recommender that covers all refactoring actions associated to the considered code smells, and (ii) putting the metric to 1 if the tool retrieved at least one recommendation, 0 otherwise.

Machine Learning Algorithms. Once we had computed dependent and independent variables, we proceeded with the definition of the machine learning algorithms to be used. In order to perform the classification, we investigated the use of multiple algorithms [16], i.e., Random Forest, Logistic Regression, Vector Space Machine, Naïve-Bayes, and Multilayer Perceptron, in order to assess what is the one giving the best performance.
from the testing strategy described above, reporting the number of
true and false positives as well as the number of true and
false negatives. We analyzed these matrices by first computing
precision, recall, and F-Measure [5]. Then, we computed the
Matthew’s Correlation Coefficient (MCC), a correlation index
between the observed and predicted binary classifications, and
the Area Under the ROC Curve (AUC-ROC), which quantifies the
ability of the model to discriminate between the different classes.

4.3 RQ2. Explaining the Proposed Approach
In the context of RQ2, we took a deeper look into the performance
of the best model coming from the previous research question.
We aimed at understanding the value of the individual metrics
selected as independent variables; this step could possibly help
us explaining why the proposed approach works (or not) when
predicting the criticality of code smells. To this aim, we employed
an information gain algorithm [73], a category of methods that can
be used to quantify the gain provided by each feature to the correct
discrimination of a dependent variable. From a formal point of view, let
M be a supervised learning model, let \( F = \{ f_1, \ldots, f_n \} \) be the set
of features composing M, i.e., the metrics reported in Table 2, an
information gain algorithm measures the difference in entropy from
before to after the set M is split on a feature \( f_i \). Roughly speaking,
the algorithm gives a measure of how much uncertainty in the
model M is reduced after splitting M on feature \( f_i \). In our work,
we implement this analysis using the Gain Ratio Feature Evaluation
algorithm [73], which is directly made available by the WEKA toolkit
[31]. It ranks the features \( f_1, \ldots, f_n \) in descending order based on the
entropy reduction provided by \( f_i \) to the decisions made by M,
i.e., it gives as output a ranked list where the more relevant features
are reported at the top. We ran the algorithm considering each code
smell independently.

To analyze the resulting rank and have statistically significant
conclusions, we finally exploited the Scott-Knott Effect Size Dif-
fERENCE (ESD) test [87]. This is an effect-size aware variation to
the original Scott-Knott test [76] that has been recommended for
software engineering research in previous studies [35, 41, 86] as it
(i) uses hierarchical cluster analysis to partition the set of treatment
means into statistically distinct groups according to their influence,
(ii) corrects the non-normal distribution of an input dataset, and (iii)
merges any two statistically distinct groups that have a negligible
effect size into one group to avoid the generation of trivial groups.
As effect size measure, the test relies on Cliff’s Delta (or \( d \)) [29]. To
compute the test, we used the publicly available implementation\(^5\)
provided by Tantithamthavorn et al. [87].

4.4 RQ3. Comparison with the state of the art
Finally, we investigated whether and to what extent the proposed
code smell prioritization approach overcomes the performance
of existing techniques. This step is paramount to understand the
novelty of our solution and how it may support developers better
than the baseline approaches.

The two closest techniques with respect to the one proposed
herein are those by Vidal et al. [93] and Arcelli Fontana and Zanoni
[25]: we set them as initial baselines for the comparison. In the
former, the authors proposed SpIRIT, a semi-automated technique
that relies on three main criteria, namely (1) stability, i.e., number
of previous changes applied on a smelly class over the number of
total changes applied on the system, (2) relevance, i.e., the relative
importance of the class within the system according to the feedback
given by a developer, and (3) modifiability scenarios, i.e., the number
of possible use cases of the application that risk to be impacted
by the presence of the smell according to the opinion of an expert.
The three criteria are then combined through a weighted average,
where the weights are assigned by the user of the tool. As the reader
might have noticed, SpIRIT explicitly requires the intervention of
an expert to be employed in practice: indeed, the technique has
been tested in an industrial case study involving a Java project
affected by a total of 47 code smells and requiring the interaction of
core developer of the subject application. For this reason, we could
not use it as a baseline for a \textit{in-vitro} assessment of our proposed
approach and we plan to perform a comparison with SpIRIT in our
future research agenda, as further explained in Section 7.

As for the technique proposed by Arcelli Fontana and Zanoni
[25], this is a machine learning-based solution that relies on 61
product metrics to predict how critical a certain code smell instance
is. This technique can be fully automated and, therefore, we could
use it as the baseline for our study and run it using the same depen-
dent variable, training, validation strategy, and dataset employed to
validate our approach. Once obtained the output from the baseline,
we compared it with ours by means of the same set of metrics used
in RQ1, i.e., precision, recall, F-Measure, MCC, and AUC-ROC.

5 RESULTS AND DISCUSSION
This section reports the results of our study, presenting each re-
search question independently.

5.1 RQ1. The Performance of our Model
In the context of RQ1, we aimed at assessing how well can we
predict the perceived criticality of code smells. Table 3 reports the
confusion matrices obtained when running the proposed approach
against our dataset of four code smell types, while Table 4 presents
the weighted average performance for each code smell. For the sake
of space limitations, we only report the results achieved with the
best classifier, i.e., Random Forest. A summary of the performance
of the other classifiers is available in our replication package [72].

In the first place, it is worth noting that the performance values
of our model are rather high and, indeed, it has an F-Measure that
ranges between 72% and 85%. This indicates that, in most of the
cases, our model can accurately classify the severity perceived
by developers. The worst case is represented by Shotgun Surgery,
where the model has an F-Measure of 72% and an AUC-ROC of 61%.
On the one hand, the former metric still indicates that the model is
able to correctly classify most of the instances of our dataset.
On the other hand, the latter suggests that the ability of separating
criticality classes may be further improved; this is also visible when
considering the confusion matrix for this smell (Table 3), where
we noticed that in 52% of cases the model classified non-severe
code smells as medium or severe cases. An example is represented
by the class security. JackrabbitAccessControlManager of the
JACKRABBIT project. This class has 164 lines of code and has been

\(^5\)Link: \url{https://github.com/klainfo/ScottKnottESD}
Table 3: RQ1 - RQ3. Confusion matrices obtained when running the proposed model against our dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Class/Smell</th>
<th>Blob</th>
<th>Complex Class</th>
<th>Spaghetti Code</th>
<th>Shotgun Surgery</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-severe</td>
<td>Medium</td>
<td>Severe</td>
<td>Non-severe</td>
<td>Medium</td>
</tr>
<tr>
<td>Our approach</td>
<td>54</td>
<td>10</td>
<td>0</td>
<td>46</td>
<td>17</td>
</tr>
<tr>
<td>Baseline</td>
<td>6</td>
<td>171</td>
<td>18</td>
<td>14</td>
<td>176</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>10</td>
<td>65</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4: RQ1 - RQ3. Weighted Average of the performance achieved by the experimented models against our dataset.

<table>
<thead>
<tr>
<th>Code smell</th>
<th>Model</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F-Meas.</th>
<th>MCC</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blob</td>
<td>Our approach</td>
<td>86%</td>
<td>85%</td>
<td>85%</td>
<td>75%</td>
<td>89%</td>
</tr>
<tr>
<td>Baseline</td>
<td>66%</td>
<td>69%</td>
<td>66%</td>
<td>44%</td>
<td>78%</td>
<td></td>
</tr>
<tr>
<td>Complex Class</td>
<td>Our approach</td>
<td>79%</td>
<td>81%</td>
<td>80%</td>
<td>71%</td>
<td>89%</td>
</tr>
<tr>
<td>Baseline</td>
<td>62%</td>
<td>63%</td>
<td>63%</td>
<td>33%</td>
<td>76%</td>
<td></td>
</tr>
<tr>
<td>Spaghetti Code</td>
<td>Our approach</td>
<td>90%</td>
<td>88%</td>
<td>89%</td>
<td>83%</td>
<td>92%</td>
</tr>
<tr>
<td>Baseline</td>
<td>83%</td>
<td>85%</td>
<td>84%</td>
<td>77%</td>
<td>89%</td>
<td></td>
</tr>
<tr>
<td>Shotgun Surgery</td>
<td>Our approach</td>
<td>74%</td>
<td>71%</td>
<td>72%</td>
<td>61%</td>
<td>78%</td>
</tr>
<tr>
<td>Baseline</td>
<td>33%</td>
<td>40%</td>
<td>35%</td>
<td>32%</td>
<td>61%</td>
<td></td>
</tr>
</tbody>
</table>

detected as smelly because every time it is changed an average of other 11 classes are also modified. Nevertheless, it has been subject to a relatively low number of changes (19) and defects (1), likely being less harmful than other instances of the smell. Analyzing the other misclassified cases, we noticed a similar trend: the model tends to misclassify instances because it is not always able to learn how to balance the information coming from the number of classes to be modified with the smelly one and the actual number of changes that involve the smelly instance. As such, we can claim that possible improvements to the classification model may concern the addition of combined metrics, e.g., the ratio between number of co-changing classes and number of previous changes of the smelly class.

As for the other code smells considered, the performance values are higher and all above 80% and 70% in terms of F-Measure and AUC-ROC. Hence, we can claim that the proposed model can be effectively adopted by developers to prioritize code smell instances. The best result is the one of Spaghetti Code (F-Measure=89%, AUC-ROC=92%): in this case, the model misclassifies only 31 cases (10% of the instances). By looking deeper at those cases, we could not find any evident property of the source code leading to those false positives. A similar discussion can be drawn when considering the Blob and Complex Class code smells. Part of our future research agenda includes the adoption of mechanisms able to better describe the functioning of the learners used for the classification, e.g., explainable AI algorithms [30].

**Finding 1.** The proposed model has an F-Measure ranging between 72% and 85%, hence being accurate in the classification of the perceived criticality of code smells. The worst case relates to Shotgun Surgery, where the model misclassifies non-severe instances because of its partial inability to take into account other process-related information like number of changes involving the smelly classes.

5.2 RQ3. Features Contributing to the Model

Table 5 reports the list of features contributing the most to the performance of the proposed model. As shown, each code smell has its own peculiarities. To classify Blob instances, the model mostly relies on structural metrics that capture complexity (RFC, WMC), cohesion (LCOM5, C3), and coupling (CBO) of the source code: basically, it means that the criticality of this code smell is given by a mix of various structural factors and cannot be described by just looking at them independently. At the same time, the number of previous defects affecting those instances (NF) as well as the workload of the committers (DSC) have a non-negligible effect. As such, on the one hand we can confirm previous findings that showed historical and socio-technical factors as relevant to manage code smells [62, 66]. On the other hand, our findings suggest that these metrics may possibly be useful for detecting code smells in the first place or even filtering the results of currently available detectors, so that they may give recommendations that are closer to the developer’s perceived criticality. Finally, the lines of code also contributes to the model, being however not the strongest factor—confirming again previous findings in the field [62, 64].

Table 5: RQ3. Information Gain of the independent variables of our approach. For space limits, only metrics providing significant contributions are reported.

<table>
<thead>
<tr>
<th>Code smell</th>
<th>Metric</th>
<th>Mean</th>
<th>SK-ESD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blob</td>
<td>RFC</td>
<td>0.65</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>LCOM5</td>
<td>0.57</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>NF</td>
<td>0.56</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>DSC</td>
<td>0.55</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>CBO</td>
<td>0.45</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>WMC</td>
<td>0.42</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>C3</td>
<td>0.35</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>LOC</td>
<td>0.34</td>
<td>41</td>
</tr>
<tr>
<td>Complex Class</td>
<td>CBO</td>
<td>0.39</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>WMC</td>
<td>0.54</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>LCOM5</td>
<td>0.54</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Read</td>
<td>0.54</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>NC</td>
<td>0.50</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>TSC</td>
<td>0.49</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>EXP</td>
<td>0.27</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>RFC</td>
<td>0.25</td>
<td>31</td>
</tr>
<tr>
<td>Spaghetti Code</td>
<td>Read</td>
<td>0.65</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>NF</td>
<td>0.57</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>C3</td>
<td>0.38</td>
<td>41</td>
</tr>
<tr>
<td>Shotgun Surgery</td>
<td>NC</td>
<td>0.32</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>LCOM5</td>
<td>0.31</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>AVG_CS</td>
<td>0.24</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Pers</td>
<td>0.17</td>
<td>22</td>
</tr>
</tbody>
</table>

When considering Complex Class, a similar discussion can be done. While the most impactful metrics concern with the structure of the code (CBO, WMC, LCOM5), other metrics seem to have a
relevant effect on the classification model. In particular, readability is the strongest factor after code metrics, indicating that developers consider comprehensibility important when prioritizing this code smell. Other relevant factors are the number of previous changes of classes (NC) and socio-technical aspects like experience and workload of the committers (EXP, DSC): again, this result confirms that the management of code smells may require additional information than structural aspects of source code [91]. Surprisingly, when considering Spaghetti Code instances we noticed that no structural factors strongly influence the classification. Readability is indeed the key factor leading developers to prioritize instances of this smell, followed by the number of previous defects affecting those classes (NF) and by the conceptual cohesion of classes (C3). Hence, it seems that developers prioritize instances of this smell that are semantically incoherent or that suffered from defects in the past. Our findings could again be used by code smell detection and filtering approaches to tune the list of recommendations to provide to developers.

Finally, the prioritization of the Shotgun Surgery smell is mainly driven by process-related factors. Not only the number of changes (NC) is the most powerful metric, but also the number of co-changing classes (AVG_CS) turned out to be relevant. Also, this is the only case in which the persistence of the smell (Pers.) appeared to impact the classification. These result seem to confirm that developers assess the severity of this code smell based on the intensity of the problem [61, 85], i.e., when number of changes or co-changing classes is high or when the problem is constantly affecting the codebase. Furthermore, the cohesion of the class (LCOM5) affects the classification, even though at a lower extent if compared to the contribution given for other code smell types.

**Finding 2.** The developer’s perceived criticality of code smells represents a multi-faceted problem that can be tackled considering a mix of metrics having different nature (e.g., structural or historical) and working at various levels of granularity (e.g., process or socio-technical aspects).

### 5.3 RQ3. Comparison with the state of the art

We compared the proposed model with a baseline. The results are reported in Tables 3 and 4, where we show confusion matrices and weighted performance values obtained when running the baseline against our dataset, respectively. Also in this case, we report the results obtained with the best classifier, that in this case was Logistic Regression—confirming the findings of the original authors [25].

In the first place, we can notice that the baseline is decently accurate and, indeed, its F-Measure values on Blob, Complex Class, and Spaghetti Code range between 63% and 84%. The exception is Shotgun Surgery (F-Measure=35%), where the baseline fails the classification in most of the cases. Despite its performance, however, the baseline never outperforms our technique. While this is especially true when considering Shotgun Surgery (-37% of F-Measure, -17% of AUC-ROC), also for the other code smells the difference is non-negligible: the F-Measure is 19%, 17%, and 5% lower than our model for Blob, Complex Class, and Spaghetti Code, respectively.

The main reason for these differences is likely imputable to the metrics employed. As shown in RQ2, structural aspects of source code can only partially contribute to the classification of the developer’s perceived criticality of code smells and, as such, the inclusion of factors covering other dimensions better fits the problem.

Of particular interest is the analysis of the results for the Spaghetti Code smell, where the baseline has the highest performance despite the fact that our findings in RQ2 reported structural aspects to be negligible. The baseline employs a variety of metrics that can capture different aspects of source code (e.g., coupling or cohesion) under different angles (e.g., by considering the lines of code with and without access methods). Some of the complexity metrics are highly correlated to readability of source code and its fault-proneness and, as such, they have the effect of “simulating” the presence of metrics like the one found to be relevant in RQ3. This claim is supported by an additional analysis in which we compute the correlation (using the Spearman’s test) between the metrics used by the baseline and those which turned out to be relevant in our previous analysis (Read., NF, and C3): we discovered that five of them (i.e., WMC-NAMM_type, NOMNAMM_type, AMW_type, CFNAMM_type, and num_final_static_attributes) are highly correlated, i.e., ρ>0.7, to at least one of the variables found in RQ2.

In conclusion, based on our findings we can claim that an approach solely based on structural metrics cannot be as accurate in the classification of the perceived criticality of code smells as a technique that includes information coming from other sources, confirming again that the problem of code smell management should be tackled in a more comprehensive manner.

**Finding 3.** The proposed model is, on average, 20% more accurate than the baseline when classifying the perceived criticality of code smells. Only in the case of Spaghetti Code the usage of multiple structural metrics can lead to results similar to those of our model.

### 6 THREATS TO VALIDITY

Some threats may have influenced our empirical study. This section discusses and overviews how we addressed them.

**Threats to construct validity.** Potential issues related to the relationship between theory and observation firstly refer to the dataset exploited in the study. Our goal was to define a developer-driven prioritization approach, so we needed to collect the developer’s perceived criticality of a set of code smells. To this aim, we followed a similar strategy as previous work [66, 79]: we monitored nine large open-source systems for 6 months and inquired the original developers as soon as they modified smelly classes following a similar strategy as previous work [66, 79]: we monitored nine large open-source systems for 6 months and inquired the original developers as soon as they modified smelly classes. In order to let them rank how harmful the involved code smells actually were. In so doing, we adopted some precautions. Firstly, we detected code smells using state-of-the-art tools [52, 62] that showed high accuracy, yet checking their output to remove false positives; in any case, we cannot exclude the presence of false negatives since these detectors have been validated on different datasets. Secondly, we asked preliminary questions on whether they perceived the presence of a design issue in the proposed class and recognized the same problem they were contacted for.
These questions aimed at ensuring that developers were really aware of the code smells they were assessing and, thus, could provide us with reliable feedback. Of course, we are aware that some of the developers might be peripheral contributors without the experience required to assess the harmfulness of code smells. To account for this aspect, we conducted a follow-up verification of the role of the subject developers within their corresponding projects: to this aim, we computed the number of commits they performed (i) over the entire change history of their projects and (ii) on the specific classes they were contacted for. As a result, we discovered that all our respondents have contributions that exceed the median number of commits made by all project’s developers both in terms of changes done over the history and on the smelly classes objects of our inquiry. In conclusion, we can argue that the dataset collection method is sound and allows a reliable analysis of the perceived code smells criticality.

Another threat in this category may be related to the granularity of the considered code smells, i.e., they are all computed at class-level. We focused on this family of smell since we aimed at characterizing the quality of classes under different angles using various metrics, with the final aim of assessing how well these metrics could be adopted to predict the perceived criticality of code smells. The selection of method-level design issues may have deserved a further, different investigation into the metrics that could capture the quality of methods.

Still related to the dataset, the perceived criticality assigned by developers when building the dataset might have been influenced by the co-occurrence of multiple code smells [1]. We mitigated this problem by presenting to developers classes affected by single code smell types among those considered in this paper, e.g., we only presented cases where a Blob did not occur with any of the other smells considered in the study. Nevertheless, we cannot exclude the presence of further design issues among those that we did not consider in the paper. As such, a larger experimentation would be desirable to corroborate our observations.

Finally, it is worth remarking that most of the independent metrics computed as well as the algorithms exploited (e.g., the machine learners) were computed by relying on well-tested, publicly available tools. This allowed us to reduce biases due to re-implementation. Their selection was based on convenience, and particularly on the skills that the authors have with them. There are a few exceptions, like the tool employed to assess the persistence and refactorability of code smells: in those cases, we relied on established guidelines (e.g., the refactorability metric has been implemented by taking into account previous research [92]). To verify the correct (re-)implementation, we also manually checked the output of our tools: we randomly verified a sample of 30 cases to assess whether (1) the number of refactoring actions identified by the tool matched the refactorings that we could identify by accessing the classes’ history, (2) the persistence of the smell was actually in line with the result given by the tool. This analysis gives us confidence of the reliability of our tooling.

**Threats to conclusion validity.** As for the relation between treatment and outcome, we relied on a set of widely-used metrics to evaluate the experimented techniques (e.g., precision, recall). Secondly, we exploited appropriate statistical tests to support our findings. As for the machine learning models experimented, the reported results may have been biased by the selected 10-fold cross-validation strategy. Previous research [87] has criticized it because the randomness with which it creates training and test data may lead to an under- or over-estimation of the real performance of a machine learning model—this is especially true in the case of classification algorithms [87]. To verify this possible bias, we conducted an additional analysis following the recommendation by Hall et al. [32]: we ran the experimented models multiple times and assessed the stability of the predictions performed. In particular, we ran 10 times 10-fold cross-validation and measured how many times the classification for a certain smell instance changed in those runs. We observed that in 93% of the cases, such predictions remained stable, thus allowing us to claim that the validation strategy did not bias the core findings of the study. As a final note, when building the machine learners, we also took into account common confounding effects like multi-collinearity and lack of hyper-parameter configurations.

**Threats to external validity.** As for the generalizability of the results, there are two main considerations. In the first place, we took into account four code smell types. On the one hand, we were somehow required to reduce the scope of the problem, given the amount of effort/time that we required to the involved developers. On the other hand, an extension targeting more code smells is still desirable and part of our future research agenda. The second point is related to the number of systems considered. We limited ourselves to nine open-source projects from two ecosystems. Yet, these systems are highly active and have been widely studied in the past by the research community, especially because they have various characteristics. Nevertheless, we are aware that our findings might not hold on to other (eco-)systems or in an industrial setting. It is our goal to replicate the paper in other contexts and corroborate the findings reported so far.

7 Conclusion

This paper presented a novel code smell prioritization approach based on the developers’ perceived criticality of code smells. We exploited several aspects related to code quality to predict the criticality of code smells, computed by collecting feedback from original developers about their perception of 1,332 code smell instances. Then, we applied several machine learning techniques to classify the code smell criticality in a three-level variable, and compared their results with a state-of-the-art tool. The results reported Random Forest to be the best machine learning algorithm with an F-measure ranging between 72% and 85%. Moreover, we found that our approach is, on average, 20% more accurate than the considered baseline when classifying the perceived criticality of code smells.

Future work includes (1) further improvements of the approach, e.g., by considering social network analysis metrics, (2) an experimentation with a larger number of code smells, and (3) an in-vivo assessment of our technique.

ACKNOWLEDGMENTS

Fabio gratefully acknowledges the help and financial support of the Swiss National Science Foundation through the SNF Project No. PZ00P2_186090 (TED).