Research Article

An Empirical Study of Software Metrics Diversity for Cross-Project Defect Prediction

Yiwen Zhong, Kun Song, ShengKai Lv, and Peng He

1School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China
2Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan 430062, China
3Hubei Engineering Research Center for Educational Informationization, Wuhan 430062, China

Correspondence should be addressed to Peng He; penghe@hubu.edu.cn

Received 4 September 2021; Revised 24 October 2021; Accepted 5 November 2021; Published 28 November 2021

1. Introduction

In software engineering, the conventional defect prediction approach trains a predictor using historical data of the target project and then uses it to predict defects in the subsequent version or release. This process is named as within-project defect prediction (WPDP). However, the cold-start problem is fatal for some new projects or inactive WPDP software projects. Hence, cross-project defect prediction (CPDP) overcomes this issue and has attracted much attention in recent years. In general, CPDP refers to predicting defects in a project using a predictor trained on historical data of other projects [1–3].

As illustrated in Figure 1, various software metrics such as static code, process, object-oriented, and network metrics have been employed for defect prediction. Several studies have also confirmed the discrepancy in the performance of WPDP with different metric sets [4, 5]. For example, Radjenovic et al. [4] highlight that object-oriented and process metrics perform better among six categories of software metrics.

As an artifact, the software can also be abstracted into a coarse-grained network structure based on the dependencies between classes, namely, a class dependency network (CDN) scheme. In CDN, each class is considered a node, and the dependencies between classes are the directed edges. From the perspective of complex networks, some researchers have verified that network metrics are better than code metrics in defect prediction [6, 7]. Additionally, deep learning for network analysis and specifically network embedding learning of a graph structure has attracted significant attention. This strategy aims to find a mapping function to transform each node into a low-dimensional representation. Such techniques involve deepwalk [8], node2vec [9], and
struc2vec [10]. Qu et al. [11] automatically learn the node representation from the CDN using network embedding for software defect prediction and achieve quite appealing results.

Recently, some researchers highlight that in addition to the features represented by a series of handcraft metrics, software programs have well-defined syntax, represented by abstracts syntax trees (ASTs), and rich semantic information hidden deep in the source code. The work of [12–14] has already demonstrated that programs’ semantic information helps characterize defects and improve defect prediction. Specifically, the semantic representation automatically learns from source code, exploiting a deep learning model that distinguishes code regions of different semantics.

Although some studies have investigated the possible benefits of including some measures such as static code and process metrics, none systematically assesses the impact of using various sets of metrics on defect prediction, especially CPDP. The information presented by different software metric sets commonly exhibits significant differences, especially in the context of cross-project. In other words, whether different software metric sets have a significantly distinct effect on the CPDP performance is still an open problem.

This paper focuses on comparative analysis and assesses the impact of using different metric sets to mitigate such problems. Furthermore, it also explores the optimal combination of various metrics in CPDP. The main contributions of this paper are summarized as follows:

1. We conduct a series of experiments and verify that the impact of different types of software metric sets on the performance of CPDP is significantly distinct, with the semantic metric being the most appealing, followed by the structural metric.

2. We find that the predictor built with the combination OSS (CK-OO, semantic, and structural metrics) performs best and achieves better performance than several state-of-the-art methods in terms of the $F$-measure metric.

The remainder of this paper is organized as follows. Section 2 reviews the related work in CPDP. Sections 3 and 4 describe the approach of our empirical study and the detailed experimental setups, respectively, while Sections 5 and 6 analyze and discuss the experimental results. Section 7 presents some threats to validity that may affect our study. Finally, Section 8 concludes this paper and presents the directions for future work.

2. Related Work

2.1. Cross-Project Defect Prediction. In recent years, the topic of CPDP has attracted considerable attention from both academia and industry. The most fundamental issues are how to pick the appropriate source projects for a target project and how to train a more accurate predictor through various strategies.

Turhan et al. [2] first utilize the nearest-neighbor filtering technique to prune irrelevant cross-project data, while Porto et al. [15] propose an instance filtering method by selecting the most similar instances from the training data set. Ryu et al. [16] suggest a method of hybrid instance selection using the nearest neighbor (HISNN). The results highlight that instances with solid local knowledge can be identified utilizing nearest-neighbors with the same class label.

To improve the performance of CPDP, Ni et al. [17] develop the FeSCH method and design three ranking strategies to choose the appropriate features. He et al. [18] study CPDP from the perspective of feature simplification.
and compare the performance between CPDP and WPDP. Li et al. [19] compare some well-known data filters and propose an HSBF (hierarchical select-based filter) method. Li et al. [20] analyze the impact of selection granularity of the training data on CPDP and propose a multigranularity selection strategy.

Additionally, Zhang et al. [21] provide an unsupervised approach entitled MT+ to determine the most suitable source project for each target project by considering the impact of various data transformations on the CPDP model. Kumar et al. [22] built a transfer learning scheme for CPDP by utilizing machine learning and identifying the best training data combination. Ryu et al. [23] develop a transfer cost-sensitive boosting method by considering distributional characteristics and the data imbalance for CPDP. They also [24] propose a multiobjective Naive Bayes learning method for CPDP by considering the class imbalance contexts. Poon et al. [25] suggest a credible theory-based Naive Bayes (CNB) classifier and establish a reweighting mechanism for CPDP between the source and target projects.

Besides, to address the heterogeneous defect data sets, He et al. [26] introduce a CPDP-IFS approach based on the distribution characteristics of both the source and target projects. Nam et al. [27] suggest an improved method, HDP, where the metric selection and matching build a defect predictor. Jing et al. [28] propose a unified metric representation for heterogeneous defect data named UMR. Yu et al. [29] present a feature matching and transfer (FMT) approach. Muddu et al. [30] tested the robustness of CPDP experimental research.

Considering CPDP, Herbold et al. [31] replicate 24 approaches proposed between 2008 and 2015 and evaluate their performance on five data sets. The authors claim that CPDP's model performance is sufficient for practical applications. Goel et al. [32] summarize independent variables, modeling techniques, performance evaluation criteria, and different approaches in building CPDP models but lack a more in-depth impact analysis.

With the extensive application of deep learning technology in various fields, its powerful feature generation ability has also been used for defect prediction [11–14]. For example, Wang et al. [13] generate the source code ASTs and automatically learn the program's hidden semantic and syntax features through a deep belief network. Li et al. [14] extract the structural information from ASTs through CNN and combine the semantic features with standard code features to improve the performance of software defect prediction. However, ASTs encapsulate only the abstract syntax structure of the source code, which cannot represent the program's execution process. Pan et al. [33] propose transforming the source code into program control flow graphs (CFG) to extract deeper semantic features from the code. Qu et al. [11] leverage a network embedding technique to automatically learn to encode the program's class dependency network structure into low-dimensional vector spaces to improve software defect prediction.

2.2. Software Metric. Software quality improvement through defect prediction has been relying on a wide variety of software metrics treated as features. To comprehend the relationship between diverse software metric sets for defect prediction, Chamoli et al. [34] analyze the performance of prediction models based on various software metrics and conclude that software metrics may indeed affect the models' defect prediction accuracy.

Mamedy et al. [35] identify that process metrics are worth collecting and improve the metric-based prediction models when data sets are collected from a wide range of software projects. Han et al. [36] combine code and process metrics as features and confirm that the predictive capabilities of using two features (BD_max and Pre-defects) are comparable to the results of using all 61 features. Öztürk et al. [37] suggest that quality metrics are superior in predicting imbalanced data sets than static code metrics. Xia et al. [38] search for the most critical software metrics and conclude that fewer than 10 metrics can better perform than utilizing 22 or more metrics.

Bluemke et al. [39] describe the process of choosing appropriate metrics for defect prediction. Accordingly, Jiarpakdee et al. [40] suggest that researchers should be aware of redundant metrics before constructing a defect prediction model to maximize their studies' internal validity. Caglayan et al. [41] conclude that the performance of different metric sets in building a defect prediction model depends on the project's characteristics and the targeted prediction level.

Mauša et al. [42] replicate the case study of deriving thresholds for software metrics using a statistical model based on logistic regression and analyze a more comprehensive set of software metrics. The results reveal that the threshold values of some metrics can be used to predict defect-prone modules effectively. Recently, Zhang et al. [43] suggest that an aggregation scheme can significantly alter correlations among metrics and correlations between metrics and the defect count through an analysis of 11 aggregation schemes using data collected from 255 open-source projects.

3. Problem and Method

3.1. Research Question. This paper defines CPDP as follows: given a source project $P_s$ and a target project $P_t$, CPDP aims to achieve the target prediction in $P_t$ using the knowledge extracted from $P_s$, where $P_s \neq P_t$. Let the source and target projects share the same feature cardinality and metric sets. The goal of CPDP is to learn a model from the selected source projects (training data) and apply it to the target project (test data). In the context examined here, project $P$, as a defect data set, contains $m$ instances represented as $P = \{I_1, I_2, \ldots, I_m\}$. An instance is $I_i = \{f_{1i}, f_{2i}, \ldots, f_{di}, L\}$, where $f_{ki}$ is the $k$–th dimension of the representation vector of instance $I_i$, and $d$ denotes the total dimensions, namely, the scale of the metrics. If instance $I$ is buggy, then $L$ is one; otherwise, $L$ is zero.
As mentioned in Section 2.2, various metrics can measure software complexity and quality in practice; therefore, a defect data set may contain multiple types of software metrics. According to the statistics, most public defect data sets contain at least two types of software metrics. For example, the commonly used AEEEM data set involves two sets of software metrics. According to the existing practice [11–13], deep learning technology can provide structural and semantic metrics. In other words, there are at least four sets of metrics.

Nevertheless, only a few research works based on these data sets have explored the impact of different metric sets and their combinations on the performance of CPDP, especially for handcrafted and automatically learned metrics. Spurred by that, this paper aims to find empirical evidence addressing the following three research questions.

RQ1: Is the impact of different metric sets on the performance of CPDP significantly distinct?
RQ2: Does CPDP based on hybrid metrics perform better?
RQ3: Does the optimized CPDP model outperform the baselines?

3.2. Approach. An effective prediction model affords more resources to be devoted to the bug-prone instances and consequently improves the quality of the latter instances. Existing CPDP models usually aim to improve the learning algorithm and make the predictor perform better. Hence, they always ignore the impact of software metrics on prediction performance. To answer the above research questions, we will consider constructing a CPDP model for two scenarios (Figure 2).

Scenario 1 considers CPDP modeling utilizing a single metric set, which is the most straightforward modeling method. For this case, we will investigate the performance of the predictor. Scenario 2 considers constructing a CPDP model based on multitype metric sets. Note that different colors in Figure 2 distinguish the types of indicator sets. The details of this scenario are provided in the following steps.

Step 1: Defect data sets are constructed for each project according to the software metric types. An instance is described as \( I = \{ f_{11}, f_{12}, \ldots, f_{1d_1}, f_{n1}, f_{n2}, \ldots, f_{nd_n}, L \} \), where \( n \) is the software metric category cardinality \((n \geq 1)\). As mentioned in Section 3.1, \( d_n \) denotes the dimension of the \( n \)-th set of software metrics, \( f \) is the metric value, and \( L \) the label.

Step 2: After collecting the defect data sets, we further determine a series of classification algorithms employed to learn the predictor, for example, Naive Bayes, logistic regression, and J48.

Step 3: The corresponding CPDP model based on the selected defect data sets and classification algorithms is constructed according to the specific scenario requirement.

4. Experimental Setup

4.1. Data Sets. We investigate our study on the public AEEEM data set [44], which involves five open-source projects. Table 1 lists the details of the five projects, where the second and third columns are the number of the defective and the clean instances, respectively. Each project refers to a process and CK-OO metrics. Each instance represents a class file and comprises software metrics and a dependent variable labeling defining if bugs exist in this class file. Table 2 presents all metric sets involved in our study.

Note that we expand the set of metrics to the existing data set, including structural and semantic metrics. The former is extracted from a class dependency network employing network embedding learning. Specifically, this paper utilizes the node2vec method [9] to map each class node to a low-dimensional vector. Regarding the semantic metrics, these adopt the method of [13]. Among the above metrics, the traditional code metrics are not listed because they have little effect on CPDP and are not applicable to the current data set.

Data imbalance is a crucial and unavoidable problem in software defect prediction. In our data set, due to the absence of defects, the number of nonbuggy samples is far more than the number of defective samples, with the imbalanced distribution seriously affecting the prediction accuracy. To overcome this problem, we balance the data sets with SMOTE. Additionally, since the scale of the numerical metric values is different in a data set, we normalize the metric values within the range of [0, 1] utilizing the z-score method.
4.2. Experimental Design. This section describes the entire experimental framework according to the previous three research questions, as illustrated in Figure 3. First, to conduct an impact analysis among all four metric sets, the CPDP experiments are conducted in the first scenario. This trial analyzes the differences of the software metric sets under a specific classifier. Then, we expand the experiments in the following scenario and compare the average prediction results of six cases involving different combination patterns. Finally, based on the optimal metric combination, we further verify the feasibility of the proposed CPDP model by challenging it against two current models. Once this process is completed, the answers to the three research questions of our study will be discussed.

4.3. Classifiers. Machine learning algorithms are widely used in defect prediction, with the classification algorithms being able to classify the defective modules correctly. This paper utilizes four typical classification algorithms as the primary learning algorithms.

(i) Logistic regression (LR): A widely used supervised classification algorithm that essentially solves a dichotomy problem. Due to its universality and practicability, several methods employ it for defect prediction [5, 18, 26].

(ii) Random forest (RF): A classifier that uses multiple trees to train and predict samples, aiming at reducing variance. RF has a better generalization and classification capability than typical decision trees [18].

(iii) Naive Bayes (NB): The simplest classifier based on Bayesian theory and independent hypothetical testing. It is widely accepted that NB outperforms other classifiers and thus is frequently used to build defect prediction models [5, 18, 23].

(iv) J48: A high-efficiency decision tree algorithm that uses the greedy technique for supervised classification, posing an appealing tool for defect prediction [18].

4.4. Evaluation Measures. To predict whether an instance (class file) is defective, we use binary classification technology. The possible results are true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The conventional classification evaluation measures include

<table>
<thead>
<tr>
<th>Category</th>
<th>Metric description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process metrics (15-dimension)</td>
<td>numberOfVersionsUntil, numberOfFixesUntil, numberOfAuthorsUntil, linesAddedUntil, avgLinesAddedUntil, maxLinesAddedUntil, maxLinesRemovedUntil, avgLinesRemovedUntil, codeChurnUntil, weightedAgeWithRespectTo</td>
</tr>
<tr>
<td>CK-OO metrics (17-dimension)</td>
<td>Coupling between object classes, Depth of Inheritance Tree, Afferent Couplings, Efferent Couplings, Lack of Cohesion in Methods, Number of Children, numberOfMethodsInherited, numberOfAttributesInherited, numberOfAttributes, numberOfLinesOfCode, numberOfMethods, numberOfPrivateAttributes, numberOfPrivateMethods, numberOfPublicAttributes, numberOfPublicMethods</td>
</tr>
<tr>
<td>Structural metrics</td>
<td>A d-dimensional space of features (d = 32)</td>
</tr>
<tr>
<td>Semantic metrics</td>
<td>A d-dimensional space of features (d = 32)</td>
</tr>
</tbody>
</table>

Table 1: AEEEM data set.

<table>
<thead>
<tr>
<th>Project</th>
<th>Number of buggy instances</th>
<th>Number of non-buggy instances</th>
<th>The total number of instances</th>
<th>Buggy ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eclipse</td>
<td>206</td>
<td>792</td>
<td>998</td>
<td>20.64</td>
</tr>
<tr>
<td>Equinox</td>
<td>129</td>
<td>196</td>
<td>325</td>
<td>39.69</td>
</tr>
<tr>
<td>Lucene</td>
<td>64</td>
<td>627</td>
<td>691</td>
<td>9.26</td>
</tr>
<tr>
<td>Mylyn</td>
<td>245</td>
<td>1,617</td>
<td>1,862</td>
<td>13.16</td>
</tr>
<tr>
<td>Pde</td>
<td>209</td>
<td>1,288</td>
<td>1,497</td>
<td>13.96</td>
</tr>
</tbody>
</table>

Table 2: Four types of software metrics will be involved in this paper.
precision, recall, and F-measure expressed as follows. Given the contradiction between precision and recall, we use F-measure to evaluate the prediction performance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Additionally, statistical tests assist in understanding whether a statistically significant difference between two results exists. This work utilizes the Wilcoxon signed-rank test to check whether the performance difference between the prediction models with different software metrics is significant. To further examine the effectiveness and follow the work of [13, 18, 26], we employ Cliff’s delta (δ) to measure the effect size of our approach. Cliff’s delta is a nonparametric effect size measurement scheme that quantifies the difference between the two approaches. Table 3 describes the meanings of various δ values [45].

### 5. Experimental Results

This section reports the experimental results aiming at answering the three research questions formulated in Section 3.1.

**RQ1:** Is the impact of different metric sets on the performance of CPDP significantly distinct?

This trial considers Scenario 1, with the corresponding results presented in Figure 4, and highlights that the F-measure values obtained using different metric sets are generally different, implying different levels of influence on CPDP. On the one hand, the prediction results of CPDP models based on different metric sets vary when the classifier is consistent. For instance, in Figure 4(a) and considering the semantic and process metrics, the medium values of the F-measure are 0.381 and 0.334, respectively, inferring that the semantic metric performs much better than the process metric under the J48 classifier. Additionally, a value of 0.381 is a relatively high index showing that the semantic metrics perform better than other metrics, whose index values are 0.334, 0.337, and 0.343.

Under different classifiers, the advantages of specific metrics are also unstable. Considering the semantic metric as an example, Figure 4(a) indicates that this metric performs best (0.381), but in Figures 4(c) and 4(d), this advantage is less obvious. Note that compared with the semantic metrics, other metrics perform the same instability (approximately 0.3–0.4), sometimes leading to more outliers.

Note: A negative δ value represents that the result of the latter metric set is better; on the contrary, the former is better.

To further distinguish the impact of different metric sets on CPDP, we evaluate the results in terms of the Wilcoxon signed-rank test (p value) and Cliff’s delta (δ) metric. In this study, we statistically analyze four types of metrics based on the null hypothesis, that is, two metric sets have the distribution of the same results. In Table 4, the Wilcoxon signed-rank test highlights no significant difference between the semantic and structural metrics and between process and CK-OO metrics, as both p values exceed 0.05. However, the differences between the two groups are statistically significant, especially between the CK-OO and semantic metrics (p value = 0.003).

In Table 4, the effect size δ between the structural and semantic metrics is small, and the metric is minimal between the CK-OO and process metrics. Considering the process and CK-OO metrics, the dominant effect size of the semantic metrics tends to be larger due to the negative δ values (−0.445 and −0.408), while the dominant effect size of structural metrics tends to be medium (δ = 0.295). Therefore, overall, the semantic metric performs best followed by the structural metric and then the CK-OO and process metrics.

In conclusion, based on the experimental results, the impact of different metric sets on the performance of CPDP is distinct, with a significant difference.

**RQ2:** Does CPDP based on hybrid metrics perform better?

To answer this research question, we construct a defect predictor using the logistic regression as described in Scenario 2. To simplify the presentation, we label the CPDP model with the Process and Semantic metric as CPDP-PS, OS (CK-OO and Semantic), SS (Structural and Semantic), OSS (CK-OO, Structural, and Semantic), and POSS (Process, CK-OO, Structural, and Semantic). Table 5 presents the prediction results of each target project in terms of F-measure values. The results indicate that CPDP-OSS performs best due to the greater F-measure values. For example, for Eclipse, the F-measure value of CPDP-OSS is 0.536. Compared to the remaining combinations, the performance is higher by 29.22%, 19.73%, 4.93%, and 9.58%, respectively. Additionally, the improvement increases to 62.62% compared to the previous CPDP involving single semantic metrics.

Additionally, for Equinox, the performance of CPDP-OSS exceeds 0.6 when using Eclipse, Lucene, and Pde as the source project. Note that compared with CPDP-PS and CPDP-OS, the performance improvement of CPDP-OSS for Mylyn is more prominent, exceeding 30%. Interestingly, for Lucene and Pde, their F-measure values of CPDP with single semantic metrics are more significant than that of CPDP.
Figure 4: The standardized boxplots of the performance of CPDP are achieved by different predictors based on different types of software metrics and classifiers in Scenario 1. From the bottom to the top of a standardized box plot: minimum, first quartile, median, third quartile, and maximum. Any data not included between the whiskers are plotted as a small circle. (a) J48, (b) LR, (c) NB, and (d) RF.

Table 4: The results of the Wilcoxon signed-rank test and Cliff’s effect size ($\delta$).

<table>
<thead>
<tr>
<th>Process vs. semantic</th>
<th>Structure vs. semantic</th>
<th>CK-OO vs. semantic</th>
<th>Sig. $p$ value (0.05)</th>
<th>$\delta$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Comparison of the CPDP performance in terms of $F$-measure.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Eclipse</th>
<th>Equinox</th>
<th>Target</th>
<th>Lucene</th>
<th>Mylyn</th>
<th>Pde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic</td>
<td>0.330 (+62.62%)</td>
<td>0.385 (+55.93%)</td>
<td>0.450 (+0.22%)</td>
<td>0.442 (+12.06%)</td>
<td>0.454 (+4.01%)</td>
<td></td>
</tr>
<tr>
<td>PS</td>
<td>0.415 (+29.22%)</td>
<td>0.528 (+13.63%)</td>
<td>0.382 (+18.06%)</td>
<td>0.379 (+30.61%)</td>
<td>0.380 (+24.21%)</td>
<td></td>
</tr>
<tr>
<td>OS</td>
<td>0.448 (+19.73%)</td>
<td>0.566 (+6.17%)</td>
<td>0.408 (+10.54%)</td>
<td>0.370 (+33.78%)</td>
<td>0.413 (+14.29%)</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>0.511 (+4.93%)</td>
<td>0.584 (+2.83%)</td>
<td>0.451 (—)</td>
<td>0.479 (+3.34%)</td>
<td>0.453 (+4.19%)</td>
<td></td>
</tr>
<tr>
<td>OSS</td>
<td><strong>0.536</strong></td>
<td><strong>0.601</strong></td>
<td><strong>0.451</strong></td>
<td><strong>0.495</strong></td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>POSS</td>
<td>0.489 (+9.58%)</td>
<td>0.583 (+3.07%)</td>
<td>0.424 (+6.37%)</td>
<td>0.472 (+4.87%)</td>
<td>0.436 (+8.26%)</td>
<td></td>
</tr>
</tbody>
</table>

Note: The percentages within the parentheses are the improvement rates of CPDP-OSS, and the boldfaced numbers represent the best results in the same context.
with hybrid metrics, except for that of CPDP-OSS. Besides, the performance of CPDP-SS is very close to that of CPDP-OSS, even on Lucene, as the F-measure values are the same. Therefore, it can be concluded that sometimes “more is not better.”

Overall, the results indicate that CPDP with CK-OO, structural, and semantic metrics can identify more buggy instances than the other combinations examined. Therefore, it is mandatory to consider the effect of hybrid metrics on CPDP.

RQ3: Does the optimized CPDP model outperform the baselines?

The previous results validate that it is still valuable to consider hybrid metrics during CPDP modeling. To evaluate the practicability and usefulness of CPDP-OSS, we built CPDP models using two existing approaches, that is, TCA+ [46] and TDSector [47], and performed experiments on the same data set. Table 6 presents the comparative results between our approach and the two baselines, where the max F-measure value per row is in bold. CPDP-OSS outperforms both baselines, as most boldfaced F-measure values, that is, 12 out of 20, and the average improvement rates of F-measure values belong to CPDP-OSS.

Compared with TCA+, Table 6 highlights that 6 out of 20 improved rates of our approach exceed 20%, while the maximum is 50.6%. Regarding TDSector, four cases pose an improvement exceeding 20%, and the maximum is 25.7%. With this evidence, the proposed CPDP-OSS approach is validated to be beneficial for improving the performance of a CPDP model.

6. Discussion

RQ1: Our experimental results suggest that the impact of various metrics sets on the performance of CPDP is distinct in terms of F-measure. Our findings indicate that the semantic metrics, on average, yield the best CPDP models, with structural metrics to follow. Meanwhile, according to Table 2, it is evident that the scale of these two metrics sets in our defect data sets is more significant, that is, a more extensive set of metrics may lead to better prediction. Thus, assisted by deep learning technology, deeper information can be automatically learned from the program.

The CK-OO metric and the process metrics are the most frequently used for defect prediction, with the authors in [4] arguing that the CK-OO metric has a good explanation and predictive power. Nevertheless, in our trials, both do not perform so well as expected. A possible explanation is a difference in the prediction context, leading to the conclusion that cross-project defect prediction is different from the traditional within-project defect prediction.

RQ2: For CPDP, the effectiveness of increasing the metrics diversity is proven, broadly consistent with some prior studies’ findings. Considering software metrics, D'Ambros et al. [43] investigate the prediction based on a single set of metrics and found that defect prediction models based on a single set of metrics are unstable. Hall et al. [5] also found that defect prediction models using a comprehensive combination of metrics perform well.

According to the experimental results, overall, using semantic and structural indicators affords a good prediction capability. One possible explanation is that when using AST to extract semantic information from code, the complexity of the source code has been achieved to a certain extent. Therefore, when the CK-OO metric is continuously considered, the improvement is limited.

RQ3: In the proposed CPDP-OSS approach, the advantage relies on the implicit diversity among software metrics. In Table 7, although the results show an overall improvement in the predictive performance of CPDP-OSS, the advantage is not apparent due to the p values exceeding 0.05 and the small effectiveness levels.

Several factors may prohibit revealing the apparent advantage of the proposed approach. Additionally, due to the limitation of the data set, only three/four types of software metrics are introduced, and we employ the most basic semantic and structural metrics learning model. Currently, some improved deep learning models have been used to solve this task.

To ease the complexity of the proposed approach, instead of using the complex and representative boosting and bagging algorithms, we utilize a simple logistic regression. Therefore, there is much room for improving our approach, and we believe that the advantages will be more evident after some improvement.

Although the advantages of the CPDP-OSS approach are not particularly obvious, it is more efficient than the two baseline competitor methods. Regarding TDSector, for each project, it requires manually calculating 76 metrics. However, after applying the classification on the semantic and structural features, we can avoid cumbersome calculations through machine learning and then reduce the metrics calculation to $17 + 15 + 2 = 34$, which significantly improves data processing efficiency. From this point of view, our experimental results still have great application value.

These two baselines are currently the most in-depth and representative baselines in our experimental research. There may be better baselines for comparison, and in future work, we will continue to follow up and compare them.

7. Threats to Validity

From this work, several meaningful results are obtained, but potential threats to the validity of our work remain.

Threats to construct validity primarily regard the software metrics used in this paper. The experimental data set employed from [42] is a public defect data set. According to the authors’ statement, inevitably, some links between the bug database and the source code repositories are missing. However, these data have been applied to numerous prior studies, and therefore, we argue that our results are credible and representative.

Threats to internal validity concern any confounding factor that may affect our results. First, we adopt the commonly used SMOTE method to preprocess the defect data sets due to the imbalanced data. As far as we know,
SMOTE-based oversampling techniques were widely adopted as the selection to handle the class imbalance problem in software defect prediction [48–51]. Although many improved sampling techniques have been proposed, it is reasonable to believe that it is feasible to use SMOTE-based oversampling technology in this paper.

Second, any feature selection method is not introduced during the CPDP modeling, and third, a simple connection method is directly used to generate the hybrid metrics in RQ2. Undoubtedly a complex fusion mechanism will result in better performance and greater calculation time. Fourth, we train the predictors for each classifier based on the default parameter settings configured by the Weka API. Hence, we are indeed aware that the results of our study would change if we use different settings of the above four factors.

Threats to statistical conclusion validity concern the relationship between the treatment and the outcome. In addition to the intuitive comparison of the prediction results in terms of $F$-measure, this paper also utilizes the Wilcoxon signed-rank test and Cliff’s delta effect size to compare the results. According to the significance criteria and effectiveness levels, the results indicate that the difference of various software metrics is distinct, and the introduction of diversity among software metrics is valuable. However, the advantage of our method is not noticeable compared to the two baseline methods, indicated by $|\delta|$ that is approximately 0.12.

Threats to external validity emphasize the generalization of the findings. Predictions in this paper are constructed on five open-source software systems. Although our experiments can be repeated with more open-source projects and developed with different software metrics, the empirical results for industrial software projects may differ from our main conclusions. We minimize this threat by selecting a data set that consists of parts of Eclipse, an open-source system with a solid industrial background.

### 8. Conclusions

This paper reports a comparative study of software metrics selection for CPDP, aiming to maximize the CPDP model’s diversity in terms of metric sets. Four types of software metrics are considered for modeling, and a series of experiments are conducted on five open-source projects. The study consists of (1) the impact analysis of different metric sets on CPDP, (2) exploration of the metrics’ combination, and (3) comparison between CPDP built with hybrid metric sets (CPDP-OSS) and two current state-of-the-art approaches.

The results indicate that the impact of different metric sets on CPDP is significantly distinct. Additionally, our trials indicate it is helpful for CPDP to increase the diversity of software metrics appropriately, and there are significant improvements between CPDP-OSS and the remaining combinations examined. The most significant improvement rate is up to 53.8%. Our results also highlight that CPDP-OSS outperforms two benchmarks, and the most considerable improvement rate is up to 50.6% and 25.7%.

---

**Table 6: Comparison between the proposed approach and two baseline methods: TCA+ and TDSelector.**

<table>
<thead>
<tr>
<th>Source/target</th>
<th>TCA+</th>
<th>TDSelector</th>
<th>CPDP-OSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eclipse/Equinox</td>
<td>0.600 (+5.9%)</td>
<td>0.514 (+23.7%)</td>
<td>0.636</td>
</tr>
<tr>
<td>Lucene/Equinox</td>
<td>0.620 (+4.4%)</td>
<td>0.664 (−2.5%)</td>
<td>0.648</td>
</tr>
<tr>
<td>Mylyn/Equinox</td>
<td>0.560 (−11.6%)</td>
<td>0.459 (+7.8%)</td>
<td>0.495</td>
</tr>
<tr>
<td>Pde/Equinox</td>
<td>0.600 (+4.0%)</td>
<td>0.610 (+2.3%)</td>
<td>0.624</td>
</tr>
<tr>
<td>Equinox/Eclipse</td>
<td>0.540 (−4.5%)</td>
<td>0.519 (−0.7%)</td>
<td>0.516</td>
</tr>
<tr>
<td>Lucene/Eclipse</td>
<td>0.560 (−4.5%)</td>
<td>0.594 (−9.9%)</td>
<td>0.535</td>
</tr>
<tr>
<td>Pde/Eclipse</td>
<td>0.430 (−27.7%)</td>
<td>0.455 (−8.7%)</td>
<td>0.549</td>
</tr>
<tr>
<td>Mylyn/Eclipse</td>
<td>0.480 (−13.5%)</td>
<td>0.510 (+5.9%)</td>
<td>0.545</td>
</tr>
<tr>
<td>Equinox/Lucene</td>
<td>0.270 (−50.6%)</td>
<td>0.337 (+20.7%)</td>
<td>0.407</td>
</tr>
<tr>
<td>Eclipse/Lucene</td>
<td>0.310 (−26.9%)</td>
<td>0.313 (−25.7%)</td>
<td>0.393</td>
</tr>
<tr>
<td>Mylyn/Lucene</td>
<td>0.250 (−32.0%)</td>
<td>0.298 (+10.7%)</td>
<td>0.330</td>
</tr>
<tr>
<td>Pde/Lucene</td>
<td>0.330 (−4.9%)</td>
<td>0.314 (+0.0%)</td>
<td>0.314</td>
</tr>
<tr>
<td>Equinox/Mylyn</td>
<td>0.230 (−24.7%)</td>
<td>0.264 (−8.6%)</td>
<td>0.287</td>
</tr>
<tr>
<td>Eclipse/Mylyn</td>
<td>0.360 (−13.1%)</td>
<td>0.303 (−3.2%)</td>
<td>0.313</td>
</tr>
<tr>
<td>Lucene/Mylyn</td>
<td>0.290 (−4.6%)</td>
<td>0.291 (+6.1%)</td>
<td>0.309</td>
</tr>
<tr>
<td>Pde/Mylyn</td>
<td>0.290 (−6.1%)</td>
<td>0.300 (−9.2%)</td>
<td>0.272</td>
</tr>
<tr>
<td>Equinox/Pde</td>
<td>0.330 (−21.2%)</td>
<td>0.367 (+9.0%)</td>
<td>0.400</td>
</tr>
<tr>
<td>Eclipse/Pde</td>
<td>0.390 (−21.2%)</td>
<td>0.391 (−0.0%)</td>
<td>0.391</td>
</tr>
<tr>
<td>Lucene/Pde</td>
<td>0.370 (−11.8%)</td>
<td>0.365 (−10.6%)</td>
<td>0.326</td>
</tr>
<tr>
<td>Mylyn/Pde</td>
<td>0.370 (−0.4%)</td>
<td>0.371 (−0.0%)</td>
<td>0.371</td>
</tr>
<tr>
<td>Average</td>
<td>0.410 (+6.0%)</td>
<td>0.412 (−5.1%)</td>
<td>0.433</td>
</tr>
</tbody>
</table>

**Table 7: The results of the Wilcoxon signed-rank test and Cliff’s effect size ($\delta$).**

<table>
<thead>
<tr>
<th></th>
<th>CPDP-OSS vs. TCA+</th>
<th>CPDP-OSS vs. TDSelector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sig. $p$ value</td>
<td>0.189</td>
<td>0.096</td>
</tr>
<tr>
<td>$\delta$ value</td>
<td>0.129</td>
<td>0.134</td>
</tr>
</tbody>
</table>
respectively. Therefore, it is meaningful to introduce the diversity of metric sets to improve the performance of CPDP.

Future work shall mainly focus on collecting more open-source projects to validate our approach’s generalization and improve the learning techniques of code semantic and structural information to provide a more effective CPDP model for defect prediction. The results of the $p$ values and the small Cliff’s delta values in the experiment show that compared with the two baselines, the effect of CPDP-OSS is not very significant. We will make improvements in future work and continue to experiment and test.

**Data Availability**

We investigate our study on the public AEEEM data set [44], which involves five open-source projects. Each project refers to process and CK-OO metrics. Each instance represents a class file and comprises software metrics and a dependent variable labeling defining if bugs exist in this class file.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Acknowledgments**

The authors greatly appreciate Dr. Jaechang Nam and Dr. Sinno Jialin Pan, the authors of reference [46], for providing us with the TCA source program and friendly teaching us how to use it. This work was supported by the National Key R&D Program of China (No. 2018YFB1003801), the National Natural Science Foundation of China (Nos. 61832014 and 61902114), the Science and Technology Innovation Program of Hubei Province under Grant (Nos. 2018ACA133 and 61902114), and the Open Foundation of Hubei Key Laboratory of Applied Mathematics (No. HBAM201901).

**References**