Pertanika Proceedings 1 (3): 9 - 12 (2025)

PERTANIKA JOURNALS

PERTANIKA PROCEEDINGS

Journal homepage: http://www.pertanika.upm.edu.my/

Cross-Project Defect Prediction Model Based on Enhanced Transfer Learning

Lian Haihua, Rodziah Atan*, Wan Nurhayati Wan Ab. Rahman and Mohd Hafeez Osman

Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

ABSTRACT

This study proposes a two-stage Cross-Project Defect Prediction (CPDP) framework to address class and feature distribution imbalance. It uses adversarial training-enhanced transfer learning to identify defective software modules. Its superiority is shown through experiments on public datasets and comparison with traditional models. The first stage involves feature extraction with two encoders and advanced preprocessing techniques. The second stage utilizes transfer learning, ensemble learning, fine-tuning, and the Synthetic Minority Oversampling Technique (SMOTE) method. Future research can expand its application and optimize the model to handle complex imbalances or incorporate other techniques to enhance predictive performance, increasing the practical potential of the CPDP model in software engineering.

Keywords: Adversarial training-GAN, cross-project software, defect patterns, loss function, prediction model, transfer learning

INTRODUCTION

In recent years, CPDP models have gained attention as manual testing in large software systems is difficult (Saeed & Saleem, 2023). They use machine learning to aid development and testing. However, traditional CPDP has low accuracy due to reliance on methods requiring identical statistical characteristics (Bala et al., 2022). Researchers have explored traditional methods to address class and feature imbalance (Wen et al., 2022; Zhao et al., 2021). This study presents a CPDP model based on adversarial training with ensemble learning, training, and fine-tuning. Contributions include enhancing transfer learning with two encoders, addressing class imbalance using SMOTE and _NNfilter, and

ARTICLE INFO

Article history:
Received: 30 April 2025
Published: 17 July 2025

DOI: https://doi.org/10.47836/pp.1.3.002

E-mail addresses:
lianhaihua59@gmail.com (Lian Haihua)
rodziah@upm.edu.my (Rodziah Atan)
wannurhayati@upm.edu.my (Wan Nurhayati Wan Ab. Rahman)
mohdhafeez@upm.edu.my (Mohd Hafeez Osman)
* Corresponding author

conducting experiments on public datasets to compare with traditional methods (Khleel & Nehéz, 2023; Szeghalmy & Fazekas, 2023).

PROBLEM STATEMENT

When addressing some problem of class imbalance, CPDP models struggle to adapt, potentially impacting model generalization performance due to the differences in scale, programming languages, and structural complexity among software projects (Govinda et al., 2023; Tang et al., 2021; Hu & Zhu, 2023). The first problem with this research is that the imbalance in the distribution of defect and non-defect samples affects the performance of CPDP models (class imbalance) (Qiao et al., 2020).

Different software projects may have different concerns, resulting in a large difference in the distribution of project features (feature distribution imbalance). For example, projects in the financial industry prioritize data security, performance and response time, while projects in the e-commerce sector focus on interface design and order processing. CPDP model performance and generalization ability might be decreased when faced with feature distribution imbalance (Hu & Zhu, 2023; Saeed & Saleem, 2023).

RESEARCH QUESTIONS

The purpose of this research is to answer the research questions as follows:

1. How can we classify class imbalance between the source and target projects in CPDP models?
2. How can feature distribution imbalance mitigation between the source project and the target project be designed in CPDP models?
3. What can the CPDP models help with risk management and decision making?

CONCLUSION

When comparing the experimental results of our research method and traditional machine learning models (Logistic Regression, SVM, Decision Tree) for defect prediction on the AEEEM dataset (EQ, ML, PDE, LC, JDT), we used the average results of different models on the AEEEM dataset. As shown in Table 1, the evaluated metrics include Accuracy, AUC

Table 1
From the comparative experimental average results

Method	Accuracy	AUC	F1 Score	Precision	Recall
SVM	0.36	0.37	0.17	0.33	0.32
Logistic regression	0.35	0.36	0.38	0.32	0.34
Decision Tree	0.43	0.43	0.48	0.42	0.35
Research	0.63	0.61	0.52	0.57	0.50

(Area Under the Curve), F1 Score, Precision, and Recall. From the average results, our proposed method achieved an Accuracy of 0.63, an AUC of 0.61, an F1 Score of 0.52, a Precision of 0.57, and a Recall of 0.50.

From the comparative experimental average results on the AEEEM dataset, it can be observed that my proposed method achieves higher predictive accuracy in cross-project defect prediction tasks compared to the three traditional methods.

The visualization results, shown in Figure 1, show that the designed method has achieved good results in cross-project defect prediction tasks. Specifically, a high recall rate indicates that actual defects can be effectively identified.

$$Recall = \frac{TP}{TP+FN} = \frac{115}{115+8} = 0.935$$

As shown in Figure 2 the ROC curve is closer to the upper left corner and the area under the precision-recall curve is larger, it indicates to a certain extent that the model has better ability to handle class imbalance and feature imbalance problems in cross-project defect prediction.

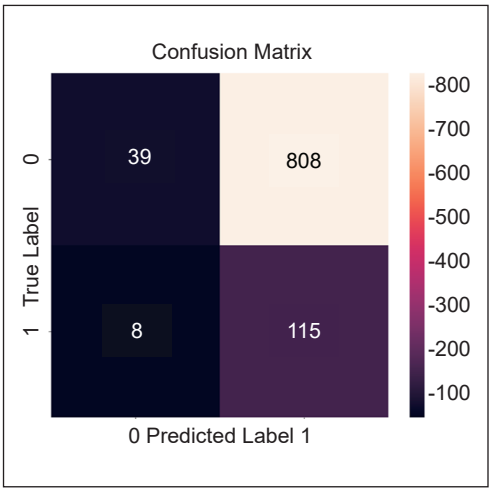

Figure 1. Result for EQ target item predicted label

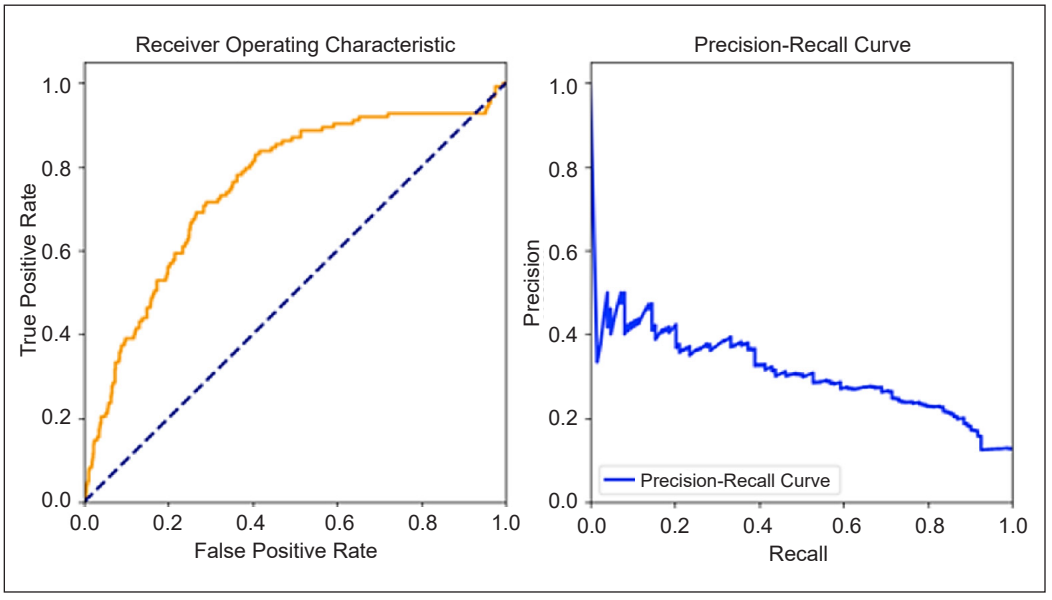

Figure 2. Result for EQ target item false positive rate and recall

ACKNOWLEDEMENT

This research owes much to many. The author is deeply grateful to Dr. Rodziah Atan for guidance, patience, and expertise. Thanks also to Dr. Wan Nurhayati Wan Ab. Rahman and Dr. Mohd Hafeez Osman for their support. The reviewers' and editors' comments improved the paper. Their encouragement kept me motivated throughout this challenging research.

REFERENCES

Bala, Y. Z., Samat, P. A., Sharif, K. Y., & Manshor, N. (2022). Improving cross-project software defect prediction method through transformation and feature selection approach. *IEEE Access*, *11*, 2318-2326. https://doi.org/10.1109/ACCESS.2022.3231456

Govinda, N. N., Lohith, R., Jha, R. K., & Gururaj, H. L. (2023). Cross-project fault prediction using artificial intelligence. *International Journal of Bioinformatics and Intelligent Computing*, *2*(1), 73-81. https://doi.org/10.61797/ijbic.v2i1.204

Hu, Z., & Zhu, Y. (2023). Cross - project defect prediction method based on genetic algorithm feature selection. *Engineering Reports*, *5*(12), Article e12670. https://doi.org/10.1002/eng2.12670

Khleel, N. A. A., & Nehéz, K. (2023). A novel approach for software defect prediction using CNN and GRU based on SMOTE Tomek method. *Journal of Intelligent Information Systems*, *60*(3), 673 - 707. https://doi.org/10.1007/s10844-023-00793-1

Qiao, L., Li, X., Umer, Q., & Guo, P. (2020). Deep learning based software defect prediction. *Neurocomputing*, *385*, 100-110. https://doi.org/10.1016/j.neucom.2019.11.067

Saeed, M. S., & Saleem, M. (2023). Cross project software defect prediction using machine learning: A review. *International Journal of Computational and Innovative Sciences*, *2*(3), 35-52.

Szeghalmy, S., & Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors, 23*(4), Article 2333. https://doi.org/10.3390/s23042333

Tang, S., Huang, S., Zheng, C., Liu, E., Zong, C., & Ding, Y. (2021). A novel cross-project software defect prediction algorithm based on transfer learning. *Tsinghua Science and Technology*, *27*(1), 41-57. https://doi.org/10.26599/TST.2020.9010040

Wen, W., Shen, C., Lu, X., Li, Z., Wang, H., Zhang, R., & Zhu, N. (2022). Cross-project software defect prediction based on class code similarity. *IEEE Access*, *10*, 105485-105495. https://doi.org/10.1109/ACCESS.2022.3211401

Zhao, Y., Zhu, Y., Yu, Q., & Chen, X. (2021). Cross-project defect prediction method based on manifold feature transformation. *Future Internet*, *13*(8), Article 216. https://doi.org/10.3390/fi13080216