


# A Framework for Homogeneous Cross-Project Defect Prediction

Lipika Goel, Amity University, India

Mayank Sharma, Amity University, India

Sunil Kumar Khatri, Amity University, India

 <https://orcid.org/0000-0003-4373-9000>

D. Damodaran, Center for Reliability, India

## ABSTRACT

Often, the prior defect data of the same project is unavailable; researchers thought whether the defect data of the other projects can be used for prediction. This made cross project defect prediction an open research issue. In this approach, the training data often suffers from class imbalance problem. Here, the work is directed on homogeneous cross-project defect prediction. A novel ensemble model that will perform in dual fold is proposed. Firstly, it will handle the class imbalance problem of the dataset. Secondly, it will perform the prediction of the target class. For handling the imbalance problem, the training dataset is divided into data frames. Each data frame will be balanced. An ensemble model using the maximum voting of all random forest classifiers is implemented. The proposed model shows better performance in comparison to the other baseline models. Wilcoxon signed rank test is performed for validation of the proposed model.

## KEYWORDS

Class Imbalance Learning, Cross Project Defect Prediction (CPDP), Ensemble Model, Homogeneous, Within Project Defect Prediction (WPDP)

## 1. INTRODUCTION

One of the vital activities and yet costly in the Software Development Process is Software Testing. It is mandatory and fundamental to manage all the limited resources the authors have in the present outline like workforce, time, monetary etc. To identify the part of software that are more likely to produce error and also requires considerations, software prediction models are useful in this scenario. Software defect prediction is one of the most heated topics at present in Software Engineering domain. Studies from the prediction models states that past data on software bugs in that particular software project can predict defects in its upcoming improvised versions. This approach is termed as Within-Project Defect Prediction (WPDP). The aspect of the training data and the machine learning techniques are used to impel and consume the conjecturing power of Software model. The WPDP examine the defect conjecture models that take up the preceding data, but the clear past records of

DOI: 10.4018/IJSI.2021010105

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

the data are maintained only by few companies. Within-Project Defect Prediction has a drawback when a project has only limited historical bug related data due to wide pertinence of Cross-Project Defect Prediction, it has been the attraction for the researchers as it reunite and collect training set of the existing models.

To solve this mentioned problem, researchers have tried to apply defect prediction in cross projects by building the models for one project and predicting the other project. This approach is known as Cross-Project Defect Prediction (CPDP). The main aim of CPDP is to predict bug-prone instances (such as classes) in a project based on the data collected from other projects. CPDP is broadly classified into Homogeneous and Heterogeneous CPDP. When the source training project has the same set of features as of target project it is known as Homogeneous CPDP whereas when the target and the source project has different metrics or features then it is termed as Heterogeneous CPDP. The feasibility and potential usefulness of CPDP built with a number of software metrics have been validated, but how to improve the performance of CPDP models is still an open issue. Through various studies it has been also concluded that suitable training data set selection can also improve the performance of the model in defect prediction. Hence training data selection from widely available public repositories is an important research area in CPDP. Since data is collected from different projects in CPDP therefore there is an imbalance in the number of defective and non defective instances. This leads to improper training of the classification model. Such models are usually biased in nature thereby impacting the performance of prediction. Figure 1 shows the difference between WPDP, Homogeneous and Heterogeneous CPDP.

The objective of this work is to propose a novel defect prediction ensemble model that will perform in a bi-fold manner. Firstly, it will handle the imbalance nature of the dataset. It will partition the training data into seven data where each data frame will have approximately equal number of defect prone and non defect prone classes. Each of the seven data frame is trained. An ensemble model based on the maximum voting of the seven Random Forest is modeled. Secondly, this proposed model will perform cross project defect prediction besides handling the class imbalance problem. 7 fold cross validation is performed to evaluate the training accuracy of the proposed model. Finally, to prove the validity of the model Wilcoxon signed rank test is performed. The research question addressed in this paper is:

RQ. Does the proposed ensemble model outperform the existing models?

The significant contributions of this paper are:

1. To develop a model to handle the class imbalance problem in CPDP.
2. To develop a standalone ensemble framework for cross project defect prediction.

The classification of the paper is as follows: Section B presents the Literature review in CPDP. Section C includes the information on the datasets, features, class imbalance problem and the target class. Section D highlights the proposed ensemble model that includes the flow from data acquisition to modeling. Section E states the performance measures that are used for evaluation of the model. Section F describes the observed and validated results. Section G concludes the paper.

## 2. STATE OF ART

Project companies do not keep a record of all the previous defects of the projects. Hence there are limitations of having trained data for making a prediction about any project. The only way is to use training data of different similar projects, and by training the model with that training data (from source project defects), predictions are made for target project defects. A survey on the same is done to give a brief summary of the literature review, to serve the researchers working in the same area.

15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/article/a-framework-for-homogeneous-cross-project-defect-prediction/266282](http://www.igi-global.com/article/a-framework-for-homogeneous-cross-project-defect-prediction/266282)

## Related Content

---

### Deep Neural Network-Based Crime Prediction Using Twitter Data

Chamith Sandagiri, Banage T. G. S. Kumaraand Banujan Kuhaneswaran (2021). *International Journal of Systems and Service-Oriented Engineering* (pp. 15-30). [www.irma-international.org/article/deep-neural-network-based-crime-prediction-using-twitter-data/272542](http://www.irma-international.org/article/deep-neural-network-based-crime-prediction-using-twitter-data/272542)

### Encapsulation of Complex HPC Services

Alexander Kipp, Ralf Schneiderand Lutz Schubert (2013). *Integrated Information and Computing Systems for Natural, Spatial, and Social Sciences* (pp. 153-176). [www.irma-international.org/chapter/encapsulation-complex-hpc-services/70608](http://www.irma-international.org/chapter/encapsulation-complex-hpc-services/70608)

### Engineering and Reengineering of Technology Enhanced Learning Scenarios Using Context Awareness Processes

Clara Inés Peña de Carrillo, Christophe Choquet, Christophe Després, Sébastien Iksal, Pierre Jacoboni, Aina Lekira, El Amine Ouraibaand Diem Pham Thi-Ngoc (2014). *Software Design and Development: Concepts, Methodologies, Tools, and Applications* (pp. 1289-1313). [www.irma-international.org/chapter/engineering-reengineering-technology-enhanced-learning/77758](http://www.irma-international.org/chapter/engineering-reengineering-technology-enhanced-learning/77758)

### Energy-Aware VM Scheduler: A Systematics Review

Ram Narayan Shuklaand Anoop Kumar Chaturvedi (2022). *International Journal of Information System Modeling and Design* (pp. 1-15). [www.irma-international.org/article/energy-aware-vm-scheduler/297631](http://www.irma-international.org/article/energy-aware-vm-scheduler/297631)

### Framework-Based Debugging for Embedded Systems

Gokhan Tanyeri, Trish Messiterand Paul Beckett (2014). *Handbook of Research on Embedded Systems Design* (pp. 424-454). [www.irma-international.org/chapter/framework-based-debugging-for-embedded-systems/116121](http://www.irma-international.org/chapter/framework-based-debugging-for-embedded-systems/116121)