

MAIN CONTRIBUTIONS OF THE DOCTORAL DISSERTATION

Thesis Topic: Software fault prediction based on source code.

Specialization: Computer Science

Code: 9.48.01.01

PhD Candidate: Hà Thị Minh Phương

Supervisors:

1. Assoc. Prof. PhD. Nguyễn Thanh Bình

2. PhD. Lê Thị Mỹ Hạnh

Institution: The University of Da Nang - Da Nang University of Science and Technology.

CONTRIBUTIONS OF THE DISSERTATION

- First, this dissertation investigates and proposes feature selection methods to eliminate redundant and irrelevant features from software defect prediction datasets. These methods reduce data noise and model complexity while improving the quality of the input data and enhancing the accuracy of software defect prediction. Furthermore, the dissertation systematically analyzes and evaluates the effectiveness of feature selection techniques and proposes their integration with data imbalance handling methods to improve prediction performance in software defect prediction tasks involving imbalanced class distributions..

- Second, this dissertation proposes and experimentally evaluates the application of Generative Adversarial Networks (GANs) and their variants to address the class imbalance problem in tabular software defect datasets. GAN-based models are employed to generate high-quality synthetic samples for the minority class, thereby balancing the class distribution and improving the predictive accuracy and generalization capability of software defect prediction models. The experimental findings provide valuable empirical evidence for selecting appropriate data generation techniques for imbalanced machine learning problems in software engineering.

Building upon the optimized feature set, the dissertation further proposes an integrated framework that combines Recursive Feature Elimination (RFE) and

Autoencoder-based feature selection with GAN-based data generation models to address data imbalance in software defect prediction. Experimental results demonstrate that integrating RFE and Autoencoder with GAN variants, particularly CTGAN, WGAN-GP, and VanillaGAN, produces higher-quality training datasets and enables defect prediction models to achieve superior performance compared with conventional oversampling techniques such as SMOTE and ADASYN. These findings confirm that jointly optimizing feature selection and data balancing constitutes an effective approach for enhancing the prediction accuracy, generalization capability, and robustness of software defect prediction models.

- Third, this dissertation proposes a novel approach that exploits both the syntactic and semantic information of source code, enabling the prediction model to learn semantically meaningful program representations and thereby improving the accuracy and generalization capability of software defect prediction. Based on this approach, the dissertation introduces a Transformer-based software defect prediction model that integrates Gensim FastText embeddings extracted from the Abstract Syntax Tree (AST) of source code. By leveraging semantic representations derived from ASTs, the proposed model effectively captures the semantic characteristics of source code tokens, leading to significant improvements in defect prediction performance for both within-project and cross-project software defect prediction tasks. Furthermore, the dissertation employs the Wilcoxon signed-rank test to assess the statistical significance of the performance differences between the proposed FastText–Transformer model and state-of-the-art baseline models. Experimental results obtained from ten Java projects demonstrate that the proposed model consistently outperforms existing software defect prediction approaches, confirming its effectiveness and superior predictive capability.

Danang, July ,2026

Supervisors' Confirmation

Supervisor

Supervisor

Doctoral Candidate

Assoc. Prof. Dr. Nguyen Thanh Binh

Dr. Le Thi My Hanh

Ha Thi Minh Phuong

