

ABSTRACT OF THE THESIS

Thesis Title: “SOFTWARE FAULT PREDICTION BASED ON SOURCECODE”

Author: Ha Thi Minh Phuong

Major: Computer Science

Code: 9480101

Training Institution: University of Science and Technology, The University of Danang

1. Objectives and Scope of Research

1.1 Objectives of the Thesis

The objective of this dissertation is to investigate the problem of software fault prediction, with a particular focus on analyzing the limitations, challenges, and unresolved issues of existing prediction approaches. Based on this analysis, the dissertation proposes and develops machine learning and deep learning techniques to enhance the performance of software fault prediction models. Specifically, the dissertation pursues the following objectives:

- The thesis conducts a comprehensive review of software fault prediction, including existing software fault prediction techniques and the current state of research and practical applications in both domestic and international contexts.
- The thesis focuses on investigating and proposing feature selection methods to eliminate redundant and irrelevant features from software fault prediction datasets. These methods reduce data noise and model complexity while improving the quality of the input data and the accuracy of software fault prediction. In addition, the thesis analyzes and evaluates the effectiveness of feature selection techniques and proposes integrating them with data imbalance handling methods to improve prediction performance on software fault datasets with imbalanced class distributions..
- Based on the optimized feature set, the thesis focuses on addressing the class imbalance problem in software fault prediction to improve the performance and accuracy of prediction models. Specifically, the thesis investigates the impact of data imbalance on software fault prediction performance and proposes integrating data sampling techniques with feature selection methods to rebalance the training dataset. In particular, the thesis proposes the use of Generative

Adversarial Networks (GANs) to generate high-quality synthetic samples that more accurately capture the underlying data distribution and the relationships among features. The integration of feature selection and data balancing techniques provides a comprehensive solution for optimizing the input data, thereby enhancing the performance, accuracy, and generalization capability of software fault prediction models..

- Motivated by the limitations of software fault prediction models based on manually engineered features, the thesis investigates a deep learning-based approach that directly exploits the syntactic and semantic information of source code. Specifically, the thesis proposes a method for leveraging the syntactic and semantic characteristics of programs, enabling the model to learn meaningful semantic representations and thereby improving the accuracy and generalization capability of software fault prediction. Based on this approach, the thesis proposes a Transformer-based software fault prediction model that incorporates Gensim FastText embeddings extracted from the Abstract Syntax Tree (AST) of source code.

1.2 Scope of Research

The research focuses on:

- Methods and techniques for software fault prediction, including feature selection techniques and data imbalance handling methods for optimizing input datasets.
- Word embedding models for learning vector representations of source code token sequences.
- Machine learning and deep learning models for developing high-performance software fault prediction systems.

2. Research Methods

The research methods include:

1. A systematic literature review of research related to software fault prediction.
2. Analytical and comparative methods to synthesize existing studies and evaluate their strengths, limitations, and research gaps.
3. A solution-oriented research methodology for proposing novel approaches based on existing studies while addressing their identified limitations.

4. An experimental methodology, including the implementation of the proposed approaches, evaluation on benchmark software fault datasets, and comparative analysis of the experimental results against state-of-the-art methods

3. Key Contributions

3.1 Scientific Contributions

- The thesis proposes an effective integration of feature selection and feature extraction techniques with advanced data sampling methods. Experimental results demonstrate that combining Recursive Feature Elimination, Autoencoder-based feature extraction, and GAN variants (including VanillaGAN, CTGAN, WGANP, CopulaGAN, and TGAN) significantly enhances the performance of traditional machine learning models in imbalanced data scenarios.
- The thesis proposes a novel FastText–Transformer model that integrates the Transformer architecture with FastText embeddings extracted from the Abstract Syntax Tree of source code. The proposed model effectively learns rich semantic representations of source code, leading to significant improvements in software fault prediction performance, particularly in both within-project and cross-project prediction scenarios.

3.2 Practical Contributions

- The thesis proposes input data optimization methods and software fault prediction models that contribute to improving fault detection during software development and testing.
- The research provides an empirical foundation for selecting appropriate feature selection, data balancing, and deep learning techniques, thereby supporting the development of software fault prediction systems with high accuracy, strong generalization capability, and robustness for real-world software projects.

4. Main Results

- The thesis proposes and evaluates the effectiveness of feature selection methods for improving the performance of software fault prediction models. In addition, it introduces an integrated approach that combines feature selection with data imbalance handling techniques to further enhance prediction performance.
- The thesis proposes an input data optimization and balancing approach by integrating feature selection methods, including Recursive Feature Elimination

and Autoencoders, with GAN-based models (including CTGAN, WGAN-GP, VanillaGAN, CopulaGAN, and TGAN) to generate synthetic samples for the minority class. Experimental results on five software fault datasets demonstrate that CTGAN, WGAN-GP, and VanillaGAN outperform traditional oversampling techniques such as SMOTE and ADASYN. The results further confirm that the integration of feature selection and GAN-based data generation significantly improves the performance and generalization capability of software fault prediction models.

- The thesis proposes, develops, and validates a novel FastText–Transformer model for software fault prediction by integrating Gensim FastText embeddings with the Transformer architecture to effectively capture the syntactic and semantic information embedded in source code through its Abstract Syntax Tree. The proposed model provides a foundation for its application in software testing support tools and automated source code quality assessment systems.

Danang, July 20, 2026

Supervisors' Confirmation

Supervisor

Supervisor

Doctoral Candidate

Assoc. Prof. Dr. Nguyen Thanh Binh

Dr. Le Thi My Hanh

Ha Thi Minh Phuong