CHAPTER I

INTRODUCTION

1.1 Background

Software defect prediction (SDP) is a well-established research domain that leverages machine learning (ML) algorithms to predict software defects (D'Ambros, Lanza, & Robbes, 2010). The goal of SDP is to estimate the likelihood of a defect occurring in a piece of software, enabling efficient resource allocation for software testing and maintenance activities (Wahono, 2015). The research community has been developing various models for SDP, each with its strengths and weaknesses. Despite significant progress, there is still ample room for improvement, particularly in the areas of prediction accuracy and handling of imbalanced data, which are the key focus areas of this research.

This research introduces an innovative approach to SDP, employing Heterogeneous Parallel Ensembles combined with Synthetic Minority Over-sampling Technique (SMOTE) and Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection.

Ensemble learning is a machine learning technique where several models are trained to solve a problem and unified to get more accurate and robust predictions (Zhang & Ma, 2012). In the context of SDP, ensembles can help improve prediction performance by aggregating the strengths of different individual models. Specifically, this research uses heterogeneous parallel ensembles, where different types of learners are trained in parallel and their outputs are combined. The use of heterogeneous learners allows the ensemble to leverage the unique strengths of each model, thus enhancing overall performance.

The selection of relevant features is critical in machine learning as it directly influences model performance. However, choosing the right features can be particularly challenging in the context of SDP due to the high-dimensionality and imbalanced nature of software defect datasets (Jureczko & Madeyski, 2010). To address these issues, this research incorporates SMOTE, a popular oversampling technique that generates synthetic samples from the minority class to balance the dataset (Chawla et al., 2002). Furthermore, LASSO, a regularization and feature selection method, is employed to select relevant features (Tibshirani, 1996). LASSO helps to avoid overfitting and improves model interpretability by generating a model that includes only the most essential features.

1.2 Problem Identification

Software defects are a significant problem in the field of software engineering, leading to increased development costs, time delays, and potential harm to the reputation of the company. While various models have been developed to predict such defects, the field still faces some considerable challenges. Some of the challenges are: Data Quality and Availability, Class Imbalance, Concept Drift, and Feature Selection. This thesis proposed to used Heterogeneous Parallel Ensembles to identify software defect before its release to production environment.

1.3 Problem Limitation

As with any research, this study on software defect prediction using heterogeneous parallel ensembles, combined with SMOTE and LASSO for feature selection, also encounters a few limitations:

1. This research only to show prediction on software defect, not the process on how to handle the issue.

2. This research only use heterogeneous parallel ensembles as method.

3. The dataset that this research use is clean MDP NASA Software Defect Dataset that can be download from this website: https://figshare.com/collections/NASA_MDP_Software_Defects_Data_Sets/4054940

1.4 Problem Definition

Based on the description of the background above, the purpose of the research is:

1. How is the performance of the use of the ensemble model in predict software defects using the clean MDP NASA Software Defect Dataset?
2. How LASSO and SMOTE can improve the performance of software defect prediction?

1.5 Research Purpose

The purpose of this research is to compare the performance of our proposed model that uses heterogeneous ensembles with LASSO and SMOTE selection using MDP NASA software defect datasets.

1.6 Outline of the Thesis.

Writing in this research is divided into at least five chapters, where each chapter has a discussion of different objectives and contents. The systematics are as follows:

Chapter 1 Introduction. This chapter contain brief introduction to Software Defect Prediction, and improvement that has been made to improve the process and the challenge in the process.

Chapter II Theoretical Background. This chapter explains theories that are used in this research related to the problem defined in Introduction Chapter. Theories covered in this chapter are Software Defect, Machine Learning using Ensemble Heterogenous Technique, Software Defect, LASSO and SMOTE. This chapter is the key to determine methodology of research in the next chapter.

Chapter III Research Methodology. This chapter contains research planning and research experiments. The research process will begin with Data Collection and Processing, Classifier Selection, Feature Selection, Data Splitting and Evaluation Metrics.

Chapter IV Result and Discussion. This chapter will explain and discuss results of experiments with commentary of what have been achieved while attaching related published paper.

Chapter V Conclusion and Suggestion. This chapter will summarize the research according to the result achieved also constructive suggestions that is potentially used for next research to achieve an even better result.