

# IMPLEMENTATION OF DATA LEVEL APPROACH TECHNIQUES TO SOLVE UNBALANCED DATA CASE ON SOFTWARE DEFECT CLASSIFICATION

Hanif Rahardian<sup>1</sup>, M. Reza Faisal<sup>2</sup>, Friska Abadi<sup>3</sup>, Radityo Adi Nugroho<sup>4</sup>,  
Rudy Herteno<sup>5</sup>

<sup>1,2,3,4,5</sup>FMIPA ULM Computer Science Study Program  
Jl. A. Yani Km 36 Banjarbaru, South Kalimantan  
Email:[rahardhyan@gmail.com](mailto:rahardhyan@gmail.com)

## Abstract

*Defects can cause significant software rework, delays, and high costs, to prevent disability it must be predictable the possibility of defects. To predict the disability the metrics software dataset is used. NASA MDP is one of the popular software metrics used to predict software defects by having 13 datasets and is generally unbalanced. The reward in the dataset can reduce the prediction of software defects because more unbalanced data produces a majority class. Data imbalance can be handled with 2 approaches, namely the data level approach technique and the algorithm level approach technique. The data level approach technique aims to improve class distribution by using resampling and data synthesis techniques. This research proposes a data level approach using resampling techniques, namely Random Oversampling (ROS), Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), Tomek Link (TL) and One-Sided Selection (OSS) which are classified with Naïve Bayes was also validated using 10 Fold Cross-Validation, then evaluated with the Area Under ROC Curve (AUC). Prediction results based on the dataset obtained the best AUC value on MC2 with a value of 0.7277 using the Synthetic Minority Oversampling Technique (SMOTE). Prediction results based on the data level approach technique obtained the best average AUC value using Tomek Link (TL) with a value of 0.62587. Prediction results based on the dataset obtained the best AUC value on MC2 with a value of 0.7277 using the Synthetic Minority Oversampling Technique (SMOTE). Prediction results based on the data level approach technique obtained the best average AUC value using Tomek Link (TL) with a value of 0.62587. Prediction results based on the dataset obtained the best AUC value on MC2 with a value of 0.7277 using the Synthetic Minority Oversampling Technique (SMOTE). Prediction results based on the data level approach technique obtained the best average AUC value using Tomek Link (TL) with a value of 0.62587.*

**Keywords:** Software defect prediction, Rewards class, Data level approach technique

## 1. INTRODUCTION

Defects are a major contributor to information technology waste and cause significant software rework, delays and high costs [2]. The highest potential defects often occur at the stage of coding the software, in order to prevent disability, it must be predictable the possibility of defects, for now the prediction of software defects focuses on estimating the number of defects in the software, finding defects and classifying defects vulnerable to defects from software components into vulnerable and non-vulnerable groups [14].

One effective method for predicting soft modules of prone to disability is to use techniques *Data Mining* which is applied to the Metrics software collected during the software development process and stored in a dataset [9]. NASA dataset (*National Aeronautics and Space Administration*) which is publicly available is a very popular software metric data in the development of software defect prediction models [7].

In general, software quality datasets are unbalanced, because generally software defects are found in a small percentage of software modules [13]. If the data are not balanced, the prediction results will tend to produce a majority class because software defects are a minority class, so many defects are not found [9].

There are two solutions to deal with class imbalance in the data that is using data level approach techniques and algorithm techniques [15]. The data level approach technique aims to improve class distribution by using resampling and data synthesis techniques [16]. Three resampling techniques to deal with class imbalances in the data are using the minor class oversampling approach, the majority class undersampling approach and the hybrid approach which are a combination of oversampling and undersampling [6].

The research conducted will deal with imbalances in the NASA MDP D dataset<sup>II</sup> using 5 resampling techniques, namely Random Oversampling (ROS), Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), Tomek Link, One-Sided Selection (OSS) and using Naïve Bayes in its classification, to find out the evaluation performance used by AUC on each resampling technique.

## 2. RESEARCH METHODS

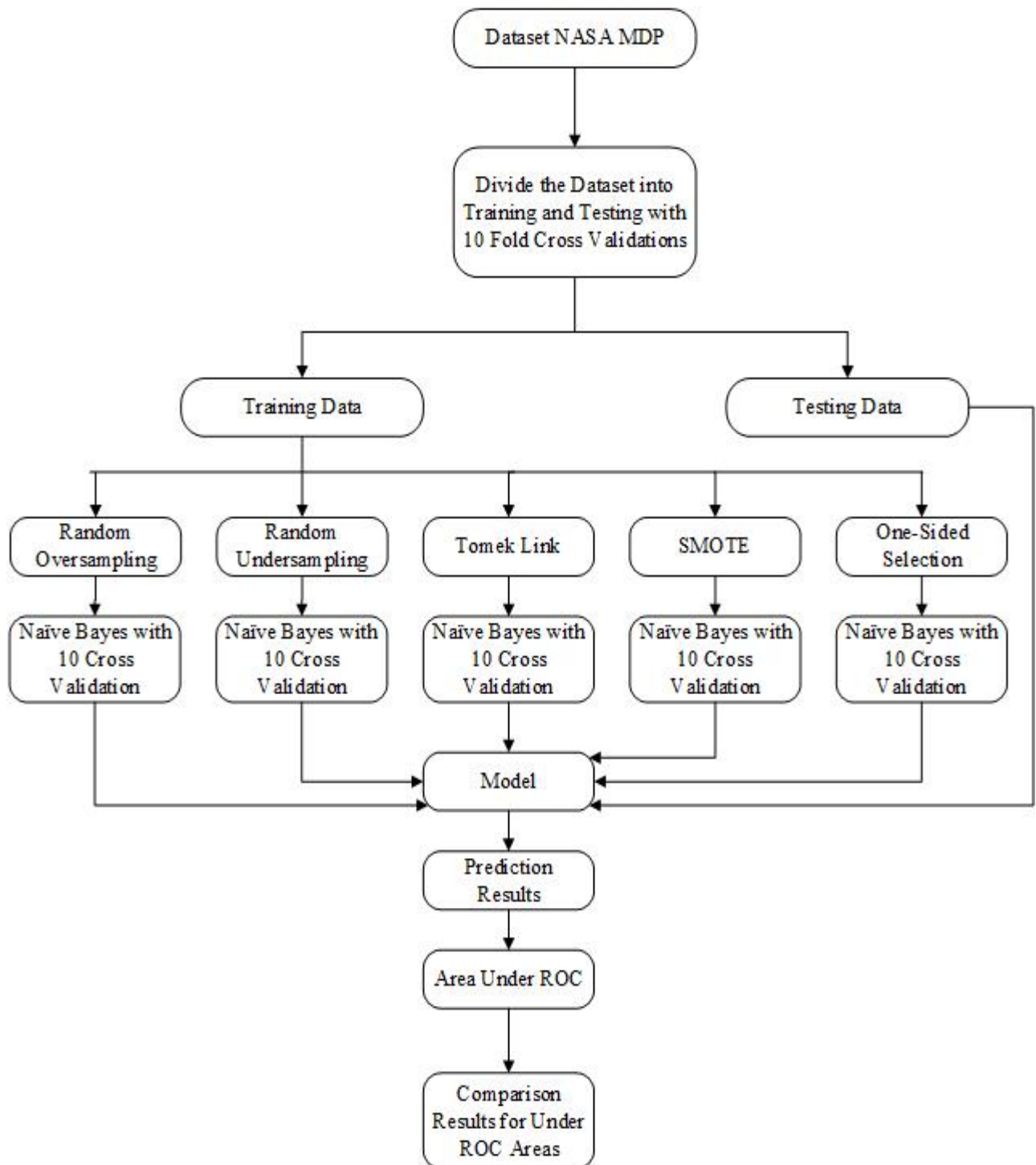


Figure 1. Research Flow

### 2.1. Data collection

In this study the dataset used was NASA MDP DII. The NASA MDP DII is the result of the original NASA MDP preprocessing, because the original NASA MDP dataset contained noise data [13]. NASA MDP DII dataset used 12, namely CM1, JM1, KC1, KC3, MC1, MC2, MW1, PC1, PC2, PC3, PC4 and PC5.

Table 1. NASA MDP DII dataset specifications

Dataset	Attribute	Amount of data	Number of Class 0	Number of Class 1
CM1	38	327	285	42
JM1	22	7782	6110	1672
KC1	22	1183	869	314
KC3	40	194	156	36
MC1	39	1998	1942	46
MC2	40	125	81	44
MW1	38	253	226	27
PC1	38	705	644	61
PC2	37	745	729	16
PC3	38	1077	943	134
PC4	38	1287	1110	177
PC5	39	1711	1240	471

## 2.2. Dividing Dataset Into Training and Testing

In this study 12 NASA MDP DII datasets will be divided into two namely training and testing with the 10 Fold Cross Validation method. The dataset will be divided into 10 data with the same value, 9 for training data and 1 for testing data. The training data will be the data used in the data level approach technique.

## 2.3. Data Level Approach Techniques

The data level approach technique will be used on 12 NASA MDP D datasets<sup>II</sup> to balance the data in the dataset. Data level approach techniques used in this study are Random Oversampling (ROS), Random Undersampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), One-Sided Selection (OSS) and Tomek Links (TL).

### 2.3.1 Random Oversampling (ROS)

*Random Oversampling* is a resampling technique that uses minority classes. In this study Random Oversampling (ROS) will add minority classes to improve and balance data with random techniques without deleting data, but adding minority classes can cause replication of the data and can cause overfitting, even though the level of accuracy is high [1].

### 2.3.2 Random Undersampling (RUS)

In this random undersampling method the majority class data will be chosen randomly, then delete or reduce the data in the majority class. This process will continue to be repeated until the amount of majority class data is equal to the amount of minority class data, by reducing class data can balance data and increase run time, but by reducing data in the class can delete important information contained in the majority data class that is deleted [5].

### 2.3.3 Synthetic Minority Oversampling Technique (SMOTE)

*Synthetic Minority Oversampling Technique* (SMOTE) is an oversampling method that handles imbalances by making samples of synthetic data, synthetic data

is created by interpolation between several examples of minority classes that are in an adjacent environment [5]. This method has limitations to some extent because the sample data is synthesized only among examples of adjacent minority classes. So this method cannot show complete data distribution [16].

#### 2.3.4 One-Sided Selection (OSS)

One-Sided Selection (OSS) is an undersampling method that only stores the most representative data from the majority class. To select the data OSS will first select one sample data  $\times$  from the majority class randomly later, using sample data from the minority class and  $\times$  as training data, OSS uses the k-Nearest Neighbors (KNN) algorithm with  $k = 1$  for classifying the remaining data from the majority class. Data that is classified correctly will be excluded from the majority class because it is considered excessive [4].

One-Sided Selection is a method that integrates with Tomek Link (TL) and Condensed Nearest Neighbor Rule (CNN). In the process Tomek Link will remove data boundaries and deviations from the majority class, then OSS will use CNN to delete some examples of majority class data that are far from the search boundary. Finally, OSS combines the minority class with the remaining majority class examples, thus the OSS creates a balanced dataset [8].

#### 2.3.5 Tomek Links (TL)

*Tomek Links* is one of the most commonly used undersampling methods, unlike other undersampling methods Tomek Link will erase most of the majority class data to produce a nearly balanced subset [11]. Tomek Link will only delete data that overlaps with other data that is labeled differently, the data will be considered as data deviation. The main idea of Tomek Link is to find overlapping data to calculate Neighbor Pairs, for example  $\times I$  and  $\times J$  if the two are close together in terms of distance, think of it as Euclidean distance, the two data are Tomek Link, if the class label of Tomek Link is different, then we can delete majority or minority class data, or both [11].

### 2.4. Data Classification

The results of balancing 12 NASA MDP DII datasets using data level approach techniques will be classified using the Naïve Bayes method. Naïve Bayes itself is the algorithm most widely used in classification problems because its simplicity, effectiveness, and robustness are very suitable for many learning scenarios, such as image classification, cheating analysis, web mining, and text classification [3]. The purpose of this classification is to predict defects that exist in software based on data that has been processed before. In the classification phase validation will also be used *Cross Validation* with the number *fold* = 10. By using *Cross Validation* can improve the accuracy of the measurement results because it reduces the likelihood of inconsistent data in the prediction stage.






### 2.5. Evaluation with Area Under ROC Curve (AUC)

Prediction results from 12 NASA MDP dataset DII will be evaluated for performance using the Area Under ROC Curve (AUC). AUC is a popular measure of performance in class imbalances, high AUC values indicate better performance [10]. AUC is calculated based on the approximate mean of trapezoidal shape fields for curves formed by TPrate and Fprate [12]. AUC can be formulated with the following equation:

$$AUC = \frac{1 + TPrate - FPrate}{2} \dots\dots\dots (1)$$

The purpose of the AUC is to get the performance results from the classification algorithm so each classification prediction result will be evaluated with AUC to find out its performance. The general guidelines used for the classification of AUC values are as follows

Table 2. General Guidelines for Area Under ROC Curve (AUC)

0.90 - 1.00	Excellent Classification	
0.80 - 0.90	Good Classification	
0.70 - 0.80	Fair Classification	
0.60 - 0.70	Poor Classification	
0.50 - 0.60	Failure	

After knowing the results of the AUC performance evaluation on 12 NASA MDP dataset DII will be compared with each other to find out the highest AUC performance value.

## 3. RESULTS AND DISCUSSION

### 3.1. Results

#### 3.1.1 Prediction Results Based on NASA MDP Dataset D<sup>II</sup>

Table 3. AUC Performance Results on 12 NASA MDP DII dataset

AUC Performance Results on 12 NASA MDP Datasets D <sup>II</sup>													
	CM1	JM1	KC1	KC3	MC1	MC2	MW1	PC1	PC2	PC3	PC4	PC5	Score flat
Original	0.5989	0.6466	.6319	0.6583	0.5348	0.7106	0.5985	0.6336	0.5151	0.5412	0.7229	.6774	0.62248
ROS	0.581	0.6473	0.6254	0.6589	0.5264	0.6947	0.5939	0.622	0.5087	0.5391	.6987	.6877	0.61532
RUS	0.5922	0.6458	0.6201	0.6508	0.5179	.705	0.601	0.5999	0.5269	0.5556	.6663	0.6674	0.61241
SMOTE	0.5871	0.6443	0.6233	0.6406	0.5265	0.7277	0.596	0.6012	0.5226	0.5208	.6713	.674	0.61128
OSS	0.6079	0.6483	0.62	0.6525	0.5343	0.6947	0.5939	0.6264	0.5048	0.541	0.7165	0.6807	0.61842
TOMEK LINK	0.6165	0.6483	0.619	0.6567	0.5322	0.68	0.68	0.6242	0.5151	0.5446	0.7144	.6794	0.62587

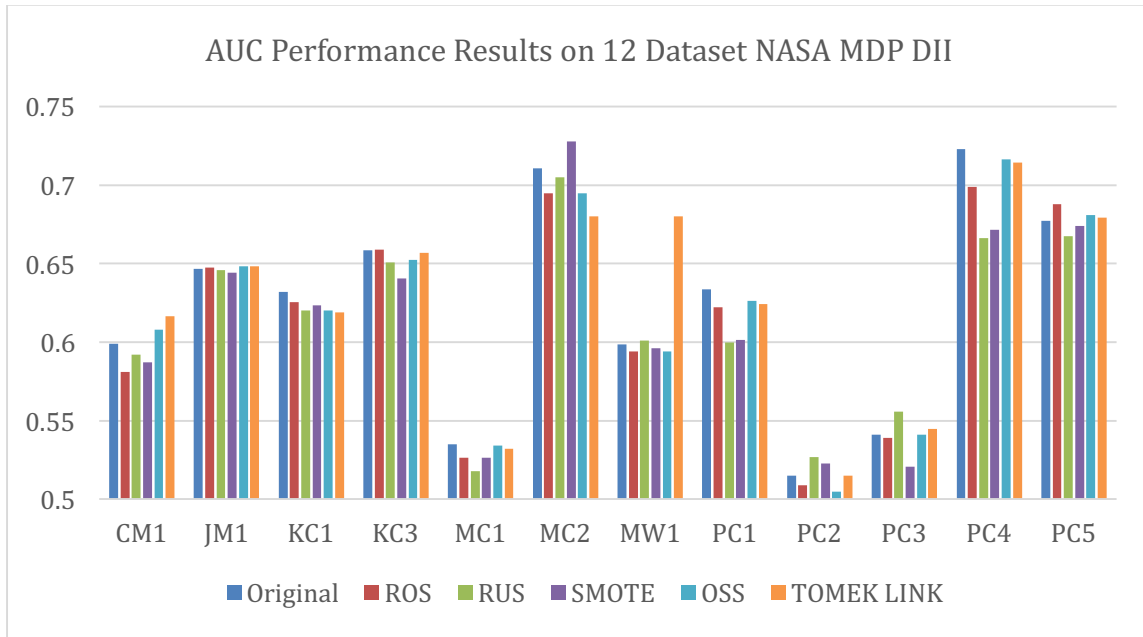


Figure 1. Graph of AUC Performance Results on 12 NASA MDP DII dataset

In this study performance evaluations have been performed using AUC level data approach techniques in each NASA MDP DII dataset. The results of the AUC performance in each dataset can be seen in table 3 and the AUC performance results graph can be seen in Figure 1. The highest AUC performance results based on the dataset are on MC2 with a performance value of 0.7277 using Synthetic Minority Oversampling Technique (SMOTE) and included in the Fair Classification while for the AUC performance value based on the lowest dataset is at PC2 of 0.5048 using the One-Sided Selection included *Failure*.

### 3.1.2 Prediction Results Based on Data Level Approach Techniques

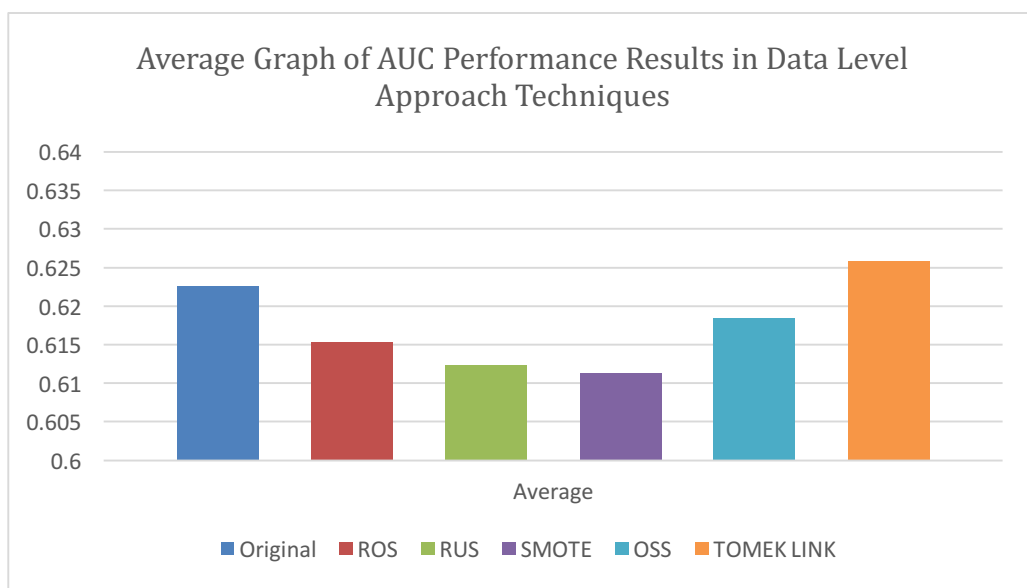


Figure 2. Average Graph of AUC Performance Results in Data Level Approach Techniques.



In this study performance evaluations have been performed using AUC level data approach techniques in each NASA MDP DII dataset. The results of the AUC performance in each dataset can be seen in table 3 and the average graph of the AUC performance in the data level approach can be seen in Figure 2.

Based on the results of the AUC performance in table 3 and the average graph of the AUC values in Figure 2 it can be concluded that the highest average value of the AUC performance results is in Tomek Link with a value of 0.62587 while the lowest average values of the AUC performance results are in Synthetic Minority Oversampling Technique (SMOTE) with a value of 0.61128.

### 3.2. DISCUSSION

In some Tomek Link datasets have high AUC performance values such as in CM1, JM1 and MW1 datasets, it can be seen from table 3 and the AUC performance results graph in Figure 1, although not always the highest but the Tomek Link AUC performance values can offset the highest AUC performance values obtained by the data level approach technique.

In this study Tomek Link does not delete too much majority class data to balance the data, compared to other data level approach techniques that are included in other undersampling methods in this study such as Random Undersampling (RUS) and One-Sided Selection (OSS). If too much deleting or reducing data in the majority class can eliminate important information in the majority data class and loss of important information in the majority data class can affect the results of predictions.

*Tomek Link* different from Random Oversampling (ROS) and Synthetic Minority Oversampling Technique (SMOTE) which are included in the oversampling method. Oversampling itself is a resampling technique that uses minority data classes, by adding or duplicating majority data classes. If too many add or duplicate the majority of data classes can cause overfitting and affect the results of predictions.

### 4. CONCLUSION

From the results and discussion of the research that has been done, it can be concluded that the results of the study that show predictions based on data level approach techniques that have the highest average AUC performance value are Tomek Link with a value of 0.62587 higher than the average value of the data level approach technique. such as Random Oversampling (ROS) with an average value of AUC performance of 0.61532, Random Undersampling (RUS) with an average value of AUC performance of 0.61241, Synthetic Minority Oversampling Technique (SMOTE) with an average value of AUC performance of 0.61128 and One-Sided Selection (OSS) with an average AUC performance value of 0.61842. Although not always the highest, the AUC Tomek Link performance value can offset the highest AUC performance value in each dataset obtained by other data level approach techniques.

Can also be seen predictions based on 12 NASA MDP dataset D<sup>II</sup> shows that the best AUC performance value on the MC2 set of 0.7277 is included in the Fair Classification using the Synthetic Minority Oversampling Technique (SMOTE) method.



## REFERENCES

- [1] Abdi, L ., & Hashemi, S. 2016. **To Combat Multi-Class Rewards Problems by Means of Over-Sampling Techniques** in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, pp. 238-251.
- [2] Anantula, PR, & Chamarthi, R. 2011. **Defect Prediction and Analysis Using ODC Approach in a Web Application**. (IJCSIT) International Journal of Computer Science and Information Technologies, 2 (5), 2242-2245.
- [3] Arar, O. Faruk & Ayan, Kursat. 2017. **A feature dependent naive bayes approach and its application to the software defect prediction problem**. Applied Computing Software.
- [4] Aridas, C. K ., Karlos, S ., Kanas, V. G ., Fazakis, N, & Kotsiantis, SB 2019. **Uncertainty Based Under Sampling for Learning Naive Bayes Classification Under Imbalanced Data Sets** in IEEE Access, vol. 8, pp. 2122-2133.
- [5] A. Saifudin, Wahono Hospital. 2015. **Data Level Approach to Handle Class Imbalance in Software Defect Prediction**. Journal of Software Engineering.
- [6] Batuwita, R., & Palade, V. 2010. **Efficient Resampling Methods for Training Support Vector Machines with Imbalanced Datasets**. Process of the International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). Barcelona: IEEE Computer Society. doi: 10.1109 / IJCNN.2010.5596787.
- [7] Hall, T ., Beecham, S ., Bowes, D ., Gray, D ., & Counsell, S. 2011. **A Systematic Literature Review on Fault Prediction Performance in Software Energizing**. IEEE Transactionson Software Engineering, Accepted for publication - available online, 1-31.
- [8] Hou, Yun ., Li, Bailin ., Li, Li ., Liu, Jiajia 2019. **A Density-based Under-sampling Algorithm for Imbalance Classification**. Journal of Physics: Conf. Series 1302 (2019) 022064.
- [9] Khoshgoftaar, Taghi M., Gao, K., & Seliya, N. 2010. **Attribute selection and imbalanced data: Problems in software defect prediction**. International conference on tools with artificial intelligence (ICTAI), vol 1, pp. 137-144.
- [10] Liu, Xu-Ying, & Zhou, Zhi-Hua. 2013. **Ensemble Methods for Class Rewards Learning. Rewards Learning: Foundations, Algorithms, and Applications**.
- [11] Luo, R. Dian., S. Wang, CC, Peng. T, Zoudong. Y., YanMei. W, Shixiong. 2018. **Bagging of Xgboost Classifiers with Random Under-sampling and Tomek Link for Noisy Label-imbalanced Data**. IOP Conf. Ser .: Mater. Sci Eng. 428 012004.
- [12] Park, BJ, Oh, SK, & Pedrycz, W. 2013. **The design of polynomial function-based Neural Network predictors for detection of software defects**. Information Sciences.
- [13] Seiffert, C., Khoshgoftaar, TM, Hulse, JV, & Folleco, A. 2011. **An Empirical Study of the Classification of Performance of Learners on Rewards and Noisy Software Quality Data**. Information Sciences, 1-25.
- [14] Song, Q., Jia, Z., Shepperd, M., Ying, S., & Liu, J. 2011. **A General Software Defect-**

- Proneness Prediction Framework.** IEEE Transactions on Software Engineering, 356-370.
- [15] Yep, BW, Rani, KA, Rahman, HA, Fong, S., Khairudin, Z., & Abdullah, NN 2014. An Application of Oversampling, ***Undersampling, Bagging and Boosting in Handling Rewarded Datasets.*** Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). 285, pp. 13-22.
- [16] Zhang, D., Liu, W., Gong, X., & Jin, H. 2011. ***A Novel Improved SMOTE Resampling Algorithm Based on Fractal.*** Computational Information Systems, 2204-2211.