THE EFFECT OF SOFTWARE METRICS ON PERFORMANCE OF SOFTWARE DEFECT CLASSIFICATION WITH ANN

Achmad Zainudin Nur [1], Mohammad Reza Faisal [2], Friska Abadi [3], Irwan Budiman [4], Rudy Herteno [5]
[12345]Ilmu Komputer FMIPA ULM
A. Yani St. KM 36 Banjarbaru, South Kalimantan
Email: J1f115202@mhs.ulm.ac.id

Abstract

Software Defect Prediction has an important role in quality software. This study uses 12 D datasets from NASA MDP which then features a selection of metrics categories software. Feature selection is performed to find out metrics software which are influential in predicting defects software. After the feature selection of the metric software category, classification will be performed using the algorithm Artificial Neural Network and validated with 5-Fold Cross Validation. Then conducted an evaluation with Area Under Curve (AUC), From datasets D" 12 NASA MDP that were evaluated with AUC, PC4, PC1 and PC3 datasets obtained the best AUC performance values. Each value is 0.915, 0.828, and 0.826 using the algorithm Artificial Neural Network.

Keywords*: Software Defect Prediction, Artificial Neural Network, Area Under Curve, NASA MDP, Cross Validation*

1. INTRODUCTION

Software Defect Prediction has an important role in the quality of software quality is Software found at the time of examination and testing. If in the examination or testing there is a defect software then it will require time and cost to repair [2]. Software Defect Prediction done to examine the performance, accuracy, precision and performance of the prediction model or the method used in research, by using a variety of datasets, such as the NASA MDP dataset.

From research conducted [5] The original or still intact NASA MDP dataset obtained through the official MDP repository website is considered to be less effective when used in predicting defects software and is considered necessary to preprocessing the dataset. That's why in the study [5] conducted preprocessing on irrelevant attributes, data redundancy and removing attributes that were considered noisy. From the results of the preprocessing that was carried out it produced the D "NASA MDP dataset.

In addition to the use of datasets, classification algorithms are also used in predicting defects software. Classification algorithms are often used to predict defects software which are used to determine which modules are included in class defects or non-defects. Classification algorithm is able to predict defects software more precisely targeted to modules that are prone to defects in the test source that

is the dataset used in a study, thereby increasing efficiency and also prediction performance [2]. As in the research conducted [1] using the Adaptive Neuro Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN), and Support Vector Machine (SVM).

Referring to the research conducted [1], in his research using three algorithms namely ANFIS, SVM and ANN and using a dataset from PROMISE (Predictor Models in Software Engineering). However, research [1] only uses the metrics category software McCabe, because it is able to carefully recognize the programming efforts but there is no explanation why not use metrics categories software other, such as the metrics category software Halstead, LoC (Line of Code), and Miscellaneous (Misc).

The research that will be conducted is to focus on using the Dataset D "NASA MDP. The feature selection will be based on metrics categories software namely, McCabe category, Halstead category, Line of Code (LoC) category, and Miscellaneous (Misc) category. And using algorithm Artificial Neural Network (ANN). To find out the results later, Area Under Curve (AUC) will be used in each category of metrics software.

2. RESEARCH METHODS

The research procedures that will be used in this study are, data collection and data processing. Flowchart to represent the research process can be seen in Figure 1

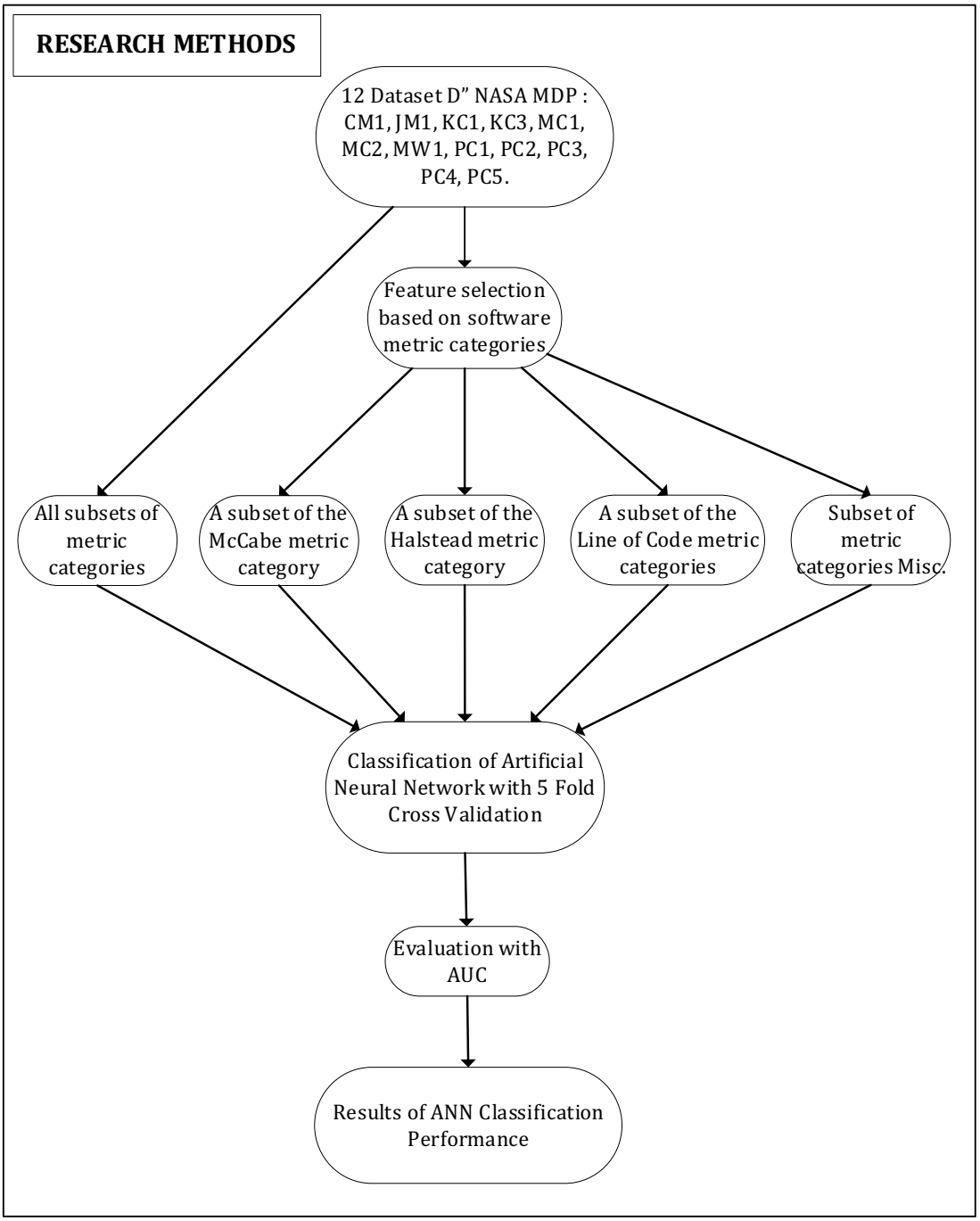

Figure 1. Research Procedure

The research procedures carried out in this study are as follows:

a. Dataset Collection The dataset
   Used in this study were 12 datasets that existed at NASA MDP as a whole. The dataset used is the dataset of research by [5], namely the dataset D "NASA MDP.

b. Feature Selection
   Selection conducted in this study is to use a selection based on metric categories software. Over all metrics software whose numbers vary in the 12 D"NASA MDP datasets are selected based on metrics software which are included in the category of metrics software namely Line of Code (LoC), Halstead, McCabe, and Miscellaneous.

c. Classification with Validation 5Fold Cross Validation
   The result of feature selection is based on the metric categories that have been selected in 12 datasets then classification is done by Algorithm Artificial Neural Network.

d. Evaluation with AUC
   Evaluate the classification results on each dataset using AUC to get the performance value of each classification algorithm. AUC was chosen as an evaluation method because according to previous research it was suitable to evaluate the value of performance or prediction performance on the dataset with problems imbalance.

   General guidelines used for the classification of [2] AUC values are as follows:
   1. 0.90 - 1.00 = Excellent Classification
   2. 0.80 - 0.90 = Good Classification
   3. 0.70 - 0.80 = Fair Classification
   4. 0.60 - 0.70 = Poor Classification
   5. 0.50 - 0.60 = Failure

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Data Collection

Dataset used in this study are 12 D "NASA MDP datasets, namely CM1, JM1, KC1, KC3, KC4, MC1, MC2, MW1, PC1, PC2, PC3, PC4, PC5 and generally consist of 40 software metrics in total.

3.1.2 Classification using ANN algorithm with 5Fold Cross Validation

The results of feature selection are based on metric categories software that have been selected in the 12 D MDAS dataset. Then do the classification with the algorithm Artificial Neural Network (ANN). And the classification algorithm is also validated using Cross Validation with the total value of Fold = 5 and also called 5Fold Cross Validation.

Cross Validation divides original data into training data and testing data. Five-fold is the definition for the value of K, where the value of K = 5. The data will be divided into a number with a specified K value of 5, it will be divided into 10 data sets. One set of data will be used as training data and the rest will be used as testing data and so on sequentially and alternately for each set [6].

3.1.3 Evaluation with Area Under Curve (AUC)

Evaluate the classification results of on each dataset using AUC to get the performance value of each classification algorithm. AUC was chosen as an evaluation method because it is suitable for evaluating performance values or predictive performance using datasets with imbalance data or unbalanced data problems. AUC is a good method used to get performance results from a classification algorithm in general and AUC is also usually used to be able to compare one classification algorithm with another [3]. AUC is a popular performance measure in the case of data class imbalance, a high AUC value indicates better performance [4].

Table 1. Results the evaluation of Artificial Neural Network with 10 Hidden Layers

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.741	0.726	0.7	0.783	0.704
JM1	0.671	0.637	0.662	0.679	0.629
KC1	0.683	0.63	0.668	0.674	0.638
KC3	0.645	0.64	0.66	0.754	0.629
MC1	0.797	0.667	0.694	0.774	0.788
MC2	0.684	0.663	0.676	0.678	0.727
MW1	0.763	0.743	0.776	0.721	0.754
PC1	0.819	0.787	0.783	0.795	0.782
PC2	0.752	0.786	0.775	0.817	0.699
PC3	0.809	0.673	0.737	0.823	0.791
PC4	0.914	0.751	0.624	0.853	0.831
PC5	0.74	0.713	0.716	0.701	0.736

Table 2. Results of the evaluation of Artificial neural Network with 8 Hidden Layers

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.738	0.727	0.701	0.782	0.703
JM1	0.67	0.638	0.661	0.679	0.629
KC1	0.681	0.629	0.667	0.672	0.638
KC3	0.636	0.642	0.66	0.759	0.63
MC1	0.798	0.668	0.694	0.746	0.788
MC2	0.666	0.663	0.674	0.675	0.742
MW1	0.76	0.743	0.774	0.721	0.754
PC1	0.825	0.788	0.785	0.793	0.783
PC2	0.749	0.786	0.773	0.817	0.701
PC3	0.813	0.673	0.74	0.822	0.788
PC4	0.915	0.754	0.624	0.847	0.833
PC5	0.738	0.712	0.716	0.707	0.732

Table 3. Results of the evaluation of Artificial neural Network with 6 Hidden Layers

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.73	0.727	0.699	0.779	0.7
JM1	0.668	0.638	0.661	0.68	0.629
KC1	0.679	0.63	0.668	0.673	0.638
KC3	0.636	0.639	0.666	0.755	0.628
MC1	0.793	0.67	0.695	0.748	0.79
MC2	0.685	0.653	0.667	0.678	0.732
MW1	0.746	0.743	0.776	0.721	0.755
PC1	0.828	0.787	0.787	0.79	0.778
PC2	0.753	0.785	0.772	0.815	0.699
PC3	0.82	0.673	0.752	0.824	0.787
PC4	0.913	0.753	0.623	0.854	0.832
PC5	0.742	0.713	0.716	0.709	0.722

Table 4. Results of the evaluation of Artificial neural Network with 4 Hidden Layers

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.718	0.728	0.698	0.782	0.697
JM1	0.668	0.64	0.66	0.68	0.629
KC1	0.684	0.63	0.669	0.672	0.638
KC3	0.643	0.639	0.661	0.756	0.663
MC1	0.783	0.669	0.694	0.75	0.79
MC2	0.67	0.66	0.667	0.676	0.759
MW1	0.752	0.743	0.776	0.723	0.753
PC1	0.817	0.788	0.789	0.788	0.773
PC2	0.754	0.777	0.772	0.817	0.701
PC3	0.815	0.673	0.756	0.825	0.791
PC4	0.914	0.751	0.622	0.85	0.835
PC5	0.746	0.712	0.715	0.709	0.735

Table 5. Results of the evaluation of Artificial neural Network with 2 Hidden Layers

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.729	0.729	0.703	0.779	0.704
JM1	0.665	0.64	0.659	0.68	0.629
KC1	0.684	0.632	0.668	0.673	0.638
KC3	0.635	0.641	0.665	0.754	0.624
MC1	0.772	0.67	0.695	0.748	0.789

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
MC2	0.715	0.663	0.669	0.679	0.736
MW1	0.742	0.737	0.752	0.721	0.752
PC1	0.81	0.789	0.79	0.789	0.771
PC2	0.755	0.747	0.772	0.819	0.696
PC3	0.818	0.675	0.766	0.826	0.792
PC4	0.901	0.747	0.619	0.842	0.836
PC5	0.732	0.714	0.716	0.706	0.724

Table 6. Results of the evaluation of Artificial neural Network with a standard value Hidden layers

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.73	0.728	0.701	0.787	0.704
JM1	0.671	0.64	0.661	0.679	0.629
KC1	0.682	0.63	0.667	0.673	0.638
KC3	0.64	0.639	0.66	0.752	0.627
MC1	0.791	0.669	0.694	0.749	0.788
MC2	0.683	0.66	0.674	0.678	0.738
MW1	0.761	0.743	0.774	0.722	0.754
PC1	0.819	0.788	0.785	0.788	0.782
PC2	0.748	0.777	0.773	0.817	0.699
PC3	0.81	0.673	0.74	0.822	0.791
PC4	0.909	0.751	0.624	0.85	0.831
PC5	0.738	0.712	0.716	0.705	0.736

Table 7. Results of best evaluation Artificial Neural Network

Dataset	Metric Categories				
	All Metrics	McCabe	Halstead	LoC	Misc
CM1	0.741	0.729	0.703	0.787	0.704
JM1	0.671	0.64	0.662	0.68	0.629
KC1	0.684	0.632	0.669	0.674	0.638
KC3	0.645	0.642	0.666	0.759	0.663
MC1	0.798	0.67	0.695	0.774	0.79
MC2	0.715	0.663	0.676	0.679	0.759
MW1	0.763	0.743	0.776	0.723	0.755
PC1	0.828	0.789	0.79	0.795	0.783
PC2	0.755	0.786	0.775	0.819	0.701
PC3	0.82	0.675	0.766	0.826	0.792
PC4	0.915	0.754	0.624	0.854	0.836
PC5	0.746	0.714	0.716	0.709	0.736

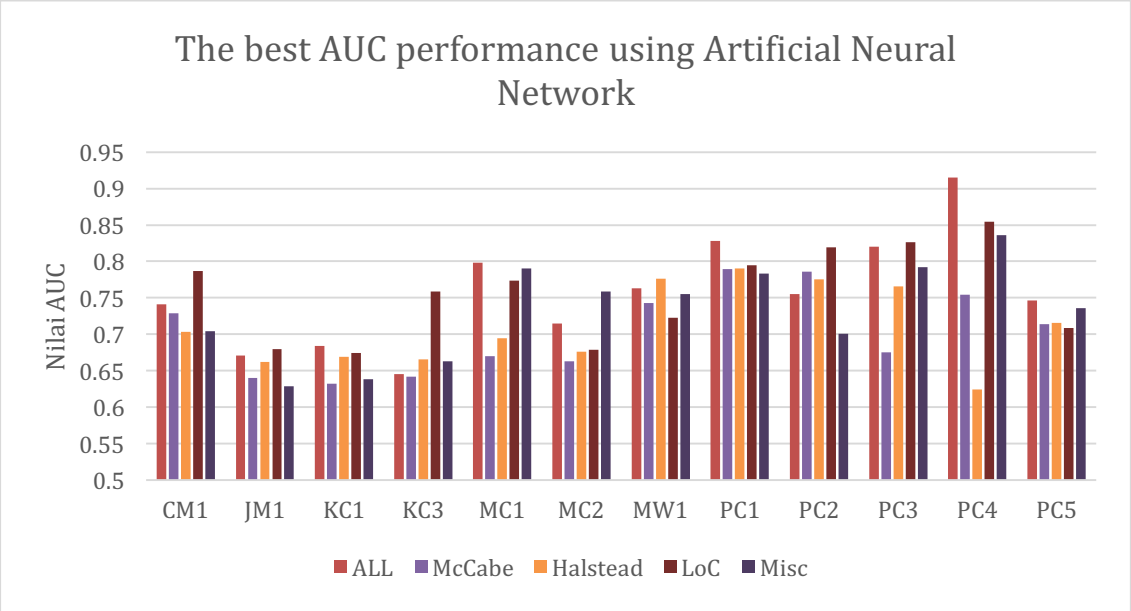

Figure 2. The best AUC performance graph using Artificial Neural Network

3.2 Discussion

AUC evaluation results without feature selection from 12 datasets using All Metrics (overall metrics) software show the best AUC performance values using the Artificial Neural Network algorithm obtained sequentially on PC4 dataset with a value of 0.915 included in Excellent Classification. PC1 dataset with a value of 0.828 is included in Good Classification and PC3 dataset with a value of 0.82 and included in Good Classification.

The results of the AUC evaluation by feature selection of 12 datasets based on the McCabe software metrics category showed that the best AUC performance values using the Artificial Neural Network algorithm were obtained sequentially on the PC1 dataset with a value of 0.789 included in the Fair Classification. PC2 dataset with a value of 0.786 is included in Fair Classification and PC4 dataset with a value of 0.754 and included in Fair Classification.

The results of the AUC evaluation by feature selection of 12 datasets based on the Halstead software metric category showed the best AUC performance values using the Artificial Neural Network algorithm were obtained sequentially on the PC1 dataset with a value of 0.79 included in Fair Classification and on the MW1 dataset with a value of 0.776 included in the Fair Classification and at PC2 dataset with a value of 0.755 is included in Fair Classification.

AUC evaluation results from software defect prediction based on the software metric category Line of Code (LoC) obtained the best AUC values sequentially using the Artificial Neural Network algorithm on PC4 dataset with a value of 0.854 included in Good Classification and on PC3 dataset with a value of 0.826 included in Good Classification and the PC2 dataset with a value of 0.819 is included in Good Classification.

The results of the AUC evaluation by feature selection of 12 datasets based on Miscellaneous software metric categories showed the best AUC performance values using the Artificial Neural Network algorithm in sequence on the PC4 dataset

with a value of 0.836 included in Good Classification and on the PC3 dataset with a value of 0792 using included in the Fair Classification and the MC1 dataset with a value of 0.79 is included in Fair Classification.

From the discussion above it can be seen that in the PC4 dataset obtained the best AUC performance value is the result without selection using all metrics software with a value of 0.915 including Excellent Classification, in the Line of Code (LoC) metric category it gets a value of 0.854 and in the Miscellaneous metric category gets a value 0.836 and both are included in Good Classification. With this it can be seen that in the PC4 dataset, the category of metrics software that affect the evaluation results and produce the best performance values are All Metrics, Line of Code and Miscellaneous.

In the PC1 dataset, the best AUC performance value was selected based on the category All Metrics (overall metrics). The value was 0.828 included in Good Classification, the category of metrics software that affected the evaluation results and produced the best performance value was All Metrics (overall metrics).

In the PC3 dataset, the best AUC performance was obtained by selecting the metric category with a Line of Code (LoC)value of 0.826 and the All Metric (overall metrics) score was 0.82 and both are included in Good Classification. With a selection based on Miscellaneous metric categories it gets a value of 0.792 and is included in the Fair Classification. With this it can be seen that in the PC3 dataset, the category of metrics software that affect the evaluation results and produce the best performance values are Line of Code and Miscellaneous.

From the discussion above it can also be seen that from 12 D MDS NASA D datasets that were evaluated with AUC, PC4, PC1, and PC3 datasets obtained the best AUC performance values. Each value is 0.915, 0.828 and 0.826 using the algorithm Artificial Neural Network.

4. CONCLUSION

The results showed that the proposed model in the Software Defect Prediction in the category All Metrics (overall metrics) obtained the best AUC value using the algorithm Artificial Neural Network on the PC4 dataset with a value of 0.915. With the McCabe category, the best AUC performance is obtained on the PC1 dataset with a value of 0.789. In the Halstead category, the best AUC performance was obtained on the PC1 dataset with a value of 0.79. With the LoC category, the best AUC performance scores were obtained on the PC4 dataset with a value of 0.854. And with the Miscellaneous category get the best AUC performance value on PC4 dataset with a value of 0.836.

It can also be seen that from 12 NASA MDP D" datasets that were evaluated with AUC, PC4, PC1 and PC3 datasets obtained the best AUC performance values. Each value is 0.915, 0.828 and 0.826 using the algorithm Artificial Neural Network.

REFERENCES

[1] Erturk, Ezgi & Ebru Akcapinar Sezer. 2015. "***A Comparison of Some Soft Computing Methods for Software Fault Prediction***." Expert Systems with Applications 42 (4): 1872–79

[2] Fitriyani & Wahono R. Satria, 2015. "***Integration of Bagging and Greedy Forward Selection in Predicting Software Defects Using Naive Bayes***". Journal of Software Engineering, Vol. 1, No. 2.

[3] Japkowicz, N. (2013). "***Assessment Metrics for Rewards Learning***". Rewards Learning: Foundations, Algorithms, and Applications.

[4] Liu, Xu-Ying, & Zhou, Zhi-Hua. 2013. "***Ensemble Methods for Class Rewards Learning***". Rewards Learning: Foundations, Algorithms, and Applications.

[5] Shepperd, M., Song, Q., Sun, M., & Mair, C. 2013. "***Data Quality: Some Comments on the NASA Software Defect Datasets***". IEEE Transactions On Software Engineering. Vol. 39, No.9.

[6] Wei, H. et al. 2019. "***Establishing a software defect prediction model via effective dimension reduction***". Information Sciences.