Metrics Based Feature Selection for Software Defect Prediction

¹Radityo Adi Nugroho, ²Friska Abadi, ³M. Reza Faisal, ⁴Rudy Herteno,
⁵Rahmat Ramadhani

^{1,2,3,4,5} Program Studi Ilmu Komputer Universitas Lambung Mangkurat
Jalan Jl. Brigjen H. Hasan Basri, Kayu Tangi, Banjarmasin

e-mail : ¹radityo.adi@ulm.ac.id, ²friska.abadi@ulm.ac.id, ³reza.faisal@ulm.ac.id, ⁴rudy.herteno@ulm.ac.id,
⁵rahmat.ramadhani@ulm.ac.id

Abstract — Nowadays, software is very influential on various sectors of life, both to solve business needs, as well as personal needs. To have a Software with high quality, testing is needed to avoid software defect. Research on software defects involving Machine Learning is currently being carried out by many researchers. This method contains one important step, which is called feature selection. In this study, researchers conducted a feature selection based on the software metric category to determine the level of accuracy of the prediction of software defects by utilizing 13 (thirteen) datasets from NASA MDP namely CM1, JM1, KC1, KC3, KC4, MC1, MC2, MW1, PC1, PC2, PC3, PC4, and PC5. To classify, the researchers involved 5 (five) classifiers, namely Naive Bayes, Decision Trees, Random Forests, K-Nearest Neighbor, and Support Vector Machines. The research result shows that each attribute on software metric categories has effect on each dataset. Naive Bayes Algorithm and Random Forest Algorithm can give better performance than other algorithm in classifieng software defect with feature selection based on metrics. On the other hand, the best metrics category on each classifier algorithm is metric Misc. From average AUC value, it can be concluded that metrics category which can give best performance is metric LoC, followed by metric Misc. Both categories have achieved highest AUC value in Random Forest classifier.

Keywords: *Software Defect Prediction, Metrics Based Feature Selection, Software Quality*

1. INTRODUCTION

Software saat ini sangat berpengaruh pada berbagai sektor kehidupan, baik untuk memenuhi kebutuhan bisnis, maupun kebutuhan pribadi. Dengan demikian perangkat lunak menjadi kebutuhan utama bagi pengguna. Oleh karena itu, kualitas software juga menjadi perhatian utama. Untuk memiliki suatu Software dengan kualitas yang tinggi maka diperlukan pengujian untuk menghindari terjadinya kerusakan software. Berbagai metode dilakukan oleh para pengembang software untuk menghindari cacat dengan melakukan berbagai metode seperti Black-box, White-box, Functional testing, Unit testing, dan lain-lain.

Pendekatan lain yang dilakukan oleh [2] dalam memprediksi kerusakan perangkat lunak adalah dengan melibatkan metode Machine Learning. Penelitian tentang cacat perangkat lunak yang melibatkan Machine Learning saat ini sedang dilakukan oleh banyak peneliti. Metode ini mengandung satu langkah penting, yang disebut pemilihan fitur. Langkah ini dapat meningkatkan kinerja klasifikasi, seperti yang dilakukan oleh banyak peneliti [3], [4], [5], dan [6].

Umumnya, penelitian tentang cacat perangkat lunak dilakukan dengan memanfaatkan kumpulan data publik seperti PROMISE dan NASA MDP [1]. Seperti yang diketahui bahwa dataset PROMISE dan NASA MDP terbentuk dari kombinasi berbagai kategori software metrics yaitu LoC (Line of Code), McCabe, Halstead dan Misc. Hal yang sedikit berbeda dilakukan oleh [7] yang melakukan penelitian dengan hanya melibatkan fitur-fitur yang masuk dalam kategori metrik software McCabe dengan alasan bahwa metrik tersebut lebih mendekati upaya pemrograman.

Pada penelitian ini peneliti melakukan pemilihan fitur berdasarkan kategori metrik perangkat lunak untuk mengetahui tingkat akurasi prediksi kecacatan perangkat lunak dengan memanfaatkan 13 (tiga belas) dataset

dari MDP NASA yaitu CM1, JM1, KC1, KC3, KC4, MC1, MC2, MW1, PC1, PC2, PC3, PC4, dan PC5. Untuk mengklasifikasikan, peneliti melibatkan 5 (lima) pengklasifikasi, yaitu Naive Bayes, Decision Trees, Random Forests, K-Nearest Neighbor, dan Support Vector Machines. Kelima algoritma tersebut dipilih karena menurut [1] merupakan algoritma klasifikasi yang paling populer dalam penelitian Software Defect Prediction.

2. STUDY LITERATUR

Penelitian tentang klasifikasi cacat perangkat lunak dengan memilih fitur untuk mendapatkan hasil yang lebih baik pada dataset MDP dan Promise NASA telah dilakukan. Pemilihan fitur yang digunakan adalah chi-squared dan korelasi pada dataset defect software yang melibatkan lima dataset MDP NASA. Hasil penelitian hanya mendapatkan sedikit fitur dari hasil seleksi fitur, yaitu sekitar 80% dari total fitur yang tersedia [3]. Selain itu, penelitian [4] dilakukan dengan menggunakan enam teknik pemilihan fitur dasar yang umum digunakan yaitu chi-square, perolehan informasi, rasio gain, dua jenis ReliefF (RF dan RFW), dan Symmetrical Uncertainty (SU). Dataset dalam penelitian ini berasal dari repositori PROMISE. Studi tersebut membandingkan pemilihan fitur dengan data sampel dan data asli. Teknik seleksi fitur juga digunakan oleh [5] dengan menggunakan Greedy Forward Selection dan Ensemble Learning Classifier yang melibatkan empat dataset MDP NASA. Studi tersebut menyatakan bahwa dengan memilih beberapa fitur berkualitas membuat hasil AUC lebih tinggi daripada yang lain. Sedangkan [6] memilih teknik pemilihan fitur Correlation-based Feature Selection (CFS) dengan melibatkan 8 (delapan) dataset MDP NASA untuk mengukur performansi dari hasil seleksi fitur yang diperoleh

Penelitian untuk meningkatkan akurasi prediksi untuk berbagai kerusakan perangkat lunak dilakukan [7] dengan hanya melibatkan metrik perangkat lunak McCabe untuk memprediksi kumpulan data dari repositori PROMISE menggunakan pengklasifikasi ANFIS.

3. METHOD

Dataset yang digunakan dalam penelitian ini diambil dari repositori MDP NASA sebanyak 13 (tiga belas) dataset, yaitu CM1, JM1, KC1, KC3, KC4, MC1, MC2, MW1, PC1, PC2, PC3, PC4, dan PC5. Setiap dataset memiliki fitur masing-masing yang terbagi menjadi 4 (empat) kriteria software metrics yaitu: LoC, Halstead, McCabe, dan Misc. Setiap atribut dan jumlah fitur pada dataset dapat dilihat pada Tabel 1 sampai dengan Tabel 5. Pada setiap dataset terdapat satu fitur yang berperan sebagai kelas yaitu fitur Defective. Fitur tersebut memiliki nilai benar atau salah. Jika suatu perangkat lunak diketahui memiliki cacat, maka itu diberikan benar, tetapi jika tidak maka itu salah.

Table 1. Attribute Detail on Metric LoC

<i>No</i>	<i>Attribute</i>
1	LOC_BLANK
2	LOC_CODE_AND_COMMENT
3	LOC_COMMENTS
4	LOC_EXECUTABLE
5	LOC_TOTAL
6	NUMBER_OF_LINES

Table 5. Attribute Detail on Metric Halstead

<i>No</i>	<i>Attribute</i>
1	HALSTEAD_CONTENT
2	HALSTEAD_DIFFICULTY
3	HALSTEAD_EFFORT
4	HALSTEAD_ERROR_EST
5	HALSTEAD_LENGTH
6	HALSTEAD_LEVEL
7	HALSTEAD_PROG_TIME
8	HALSTEAD_VOLUME
9	NUM_OPERANDS
10	NUM_OPERATORS
11	NUM_UNIQUE_OPERANDS
12	NUM_UNIQUE_OPERATORS

Table 6. Attribute Detail on Metric McCabe

<i>No</i>	<i>Attribute</i>
1	CYCLOMATIC_COMPLEXITY
2	DESIGN_COMPLEXITY
3	ESSENTIAL_COMPLEXITY
4	CYCLOMATIC_DENSITY
5	PATHOLOGICAL_COMPLEXITY

Table 7. Attribute Detail on Metric Misc

<i>No</i>	<i>Attribute</i>
1	BRANCH_COUNT
2	CALL_PAIRS
3	CONDITION_COUNT
4	DECISION_COUNT
5	DECISION_DENSITY
6	DESIGN_DENSITY
7	EDGE_COUNT
8	ESSENTIAL_DENSITY
9	GLOBAL_DATA_COMPLEXITY
10	GLOBAL_DATA_DENSITY
11	MAINTENANCE_SEVERITY
12	MODIFIED_CONDITION_COUNT
13	MULTIPLE_CONDITION_COUNT
14	NODE_COUNT
15	NORMALIZED_CYLOMATIC_COMPLEXITY
16	PARAMETER_COUNT
17	PERCENT_COMMENTS

Table 8. Number of Attribute on Dataset

<i>Dataset</i>	<i>CM1</i>	<i>JM1</i>	<i>KC1</i>	<i>KC3</i>	<i>KC4</i>	<i>MC1</i>	<i>MC2</i>	<i>MW1</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>
Attribut Number	38	22	22	40	40	39	40	38	38	37	38	38	39

Tabel 5 menunjukkan bahwa setiap dataset pada MDP NASA memiliki jumlah fitur yang berbeda. Tabel 6 menunjukkan detail perbedaan fitur pada setiap dataset.

Table 9. Feature Deatil on each NASA MDP Dataset

Metrics Category	Features	Original Dataset												
		CM 1	JM 1	KC 1	KC 3	KC 4	MC 1	MC 2	MW 1	PC 1	PC 2	PC 3	PC 4	PC 5
LoC	LOC_BLANK	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	LOC_CODE_AND_COMMENT	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	LOC_COMMENTS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	LOC_EXECUTABLE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	LOC_TOTAL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NUMBER_OF_LINES	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Halstead	HALSTEAD_CONTENT	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_DIFFICULTY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_EFFORT	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_ERROR_EST	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_LENGTH	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_LEVEL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_PROG_TIME	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	HALSTEAD_VOLUME	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NUM_OPERANDS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NUM_OPERATORS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NUM_UNIQUE_OPERANDS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NUM_UNIQUE_OPERATORS	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
McCabe	CYCLOMATIC_COMPLEXITY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	DESIGN_COMPLEXITY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	ESSENTIAL_COMPLEXITY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	CYCLOMATIC_DENSITY	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	PATHOLOGICAL_COMPLEXITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Misc	BRANCH_COUNT	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	CALL_PAIRS	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	CONDITION_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Metrics Category	Features	Original Dataset												
		CM 1	JM 1	KC 1	KC 3	KC 4	MC 1	MC 2	MW 1	PC 1	PC 2	PC 3	PC 4	PC 5
	DECISION_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	DECISION_DENSITY	Y			Y	Y		Y	Y	Y	Y	Y	Y	
	DESIGN_DENSITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	EDGE_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	ESSENTIAL_DENSITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	GLOBAL_DATA_COMPLEXITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	GLOBAL_DATA_DENSITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	MAINTENANCE_SEVERITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	MODIFIED_CONDITION_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	MULTIPLE_CONDITION_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NODE_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	NORMALIZED_CYLOMATIC_COMPLEXITY	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	PARAMETER_COUNT	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	PERCENT_COMMENTS	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Picture 7. Research Flow

Alur penelitian ini dijelaskan sebagai berikut:

- Fitur dari setiap dataset dipilih berdasarkan metrik perangkat lunak. Maka dari itu telah terbentuk dataset baru yaitu: PC5_LoC, PC5_Halstead, PC5_McCabe, PC5_Misc, MW1_LoC, MW1_Halstead, MW1_McCabe, MW1_Misc, dll.
- Dari semua dataset baru tersebut, dihasilkan model baru menggunakan algoritma klasifikasi Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbor, dan Support Vector Machine.
- Setelah model siap, dilakukan evaluasi dengan menggunakan 10 Fold Cross Validation dan keakuratannya dihitung menggunakan nilai AUC.
- Nilai ABK dari setiap pengklasifikasi pada setiap kategori metrik dirata-ratakan untuk mendapatkan nilai ABK dari setiap kategori metrik..

4. RESULT

3.3 Naive Bayes Classifier

Hasil klasifikasi setiap metrik pada setiap dataset dengan menggunakan Naive Bayes dapat dilihat dari Tabel 7. Dari rata-rata AUC pada Tabel 8 terlihat bahwa nilai klasifikasi terbaik pada algoritma Naive Bayes adalah metric Misc.

Table 10. AUC Value on Naive Bayes

<i>Metrics</i>	<i>AUC Value</i>												
	<i>CM1</i>	<i>JM1</i>	<i>KC1</i>	<i>KC3</i>	<i>KC4</i>	<i>MC1</i>	<i>MC2</i>	<i>MW1</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>
LoC	0,773	0,654	0,777	0,841	0,455	0,86	0,602	0,744	0,729	0,851	0,808	0,804	0,93
Halstead	0,739	0,68	0,79	0,813	0,5	0,857	0,692	0,774	0,663	0,88	0,71	0,677	0,928
McCabe	0,758	0,594	0,728	0,768	0,692	0,89	0,712	0,683	0,718	0,789	0,653	0,738	0,938
Misc	0,71	0,585	0,733	0,812	0,762	0,926	0,721	0,719	0,713	0,84	0,714	0,822	0,943

Table 11. Average AUC value on Naive Bayes

<i>Metrics</i>	<i>Average AUC</i>
LoC	0,756
Halstead	0,746
McCabe	0,743
Misc	0,769

3.4 Decision Tree Classifier

Hasil klasifikasi masing-masing metrik pada setiap dataset menggunakan Decision Tree dapat dilihat pada Tabel 9. Dari rata-rata AUC pada Tabel 10 terlihat bahwa nilai klasifikasi terbaik pada Decision Tree adalah metric Misc.

Table 12. AUC Value on Decision Tree

<i>Metrics</i>	<i>AUC Value</i>												
	<i>CM1</i>	<i>JM1</i>	<i>KC1</i>	<i>KC3</i>	<i>KC4</i>	<i>MC1</i>	<i>MC2</i>	<i>MW1</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>
LoC	0,494	0,506	0,5	0,549	0,56	0,5	0,505	0,487	0,561	0,525	0,504	0,622	0,532
Halstead	0,488	0,505	0,529	0,484	0,5	0,5	0,548	0,519	0,496	0,525	0,5	0,504	0,762
McCabe	0,498	0,509	0,5	0,535	0,683	0,5	0,516	0,564	0,5	0,5	0,5	0,5	0,668
Misc	0,515	0,506	0,5	0,473	0,732	0,493	0,585	0,524	0,5	0,525	0,503	0,754	0,754

Table 13. Average AUC value on Decision Tree

<i>Metrics</i>	<i>Average AUC</i>
LoC	0,527
Halstead	0,528
McCabe	0,536
Misc	0,566

3.5 Random Forest Classifier

Hasil klasifikasi setiap metrik pada setiap dataset dengan menggunakan Random Forest dapat dilihat dari Tabel 11. Dari rata-rata AUC pada Tabel 12 terlihat bahwa nilai klasifikasi terbaik pada Random Forest adalah metrik LOC.

Table 14. AUC Value on Random Forest

<i>Metrics</i>	<i>AUC Value</i>												
	<i>CMI</i>	<i>JMI</i>	<i>KCI</i>	<i>KC3</i>	<i>KC4</i>	<i>MC1</i>	<i>MC2</i>	<i>MW1</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>
LoC	0,755	0,702	0,796	0,816	0,696	0,968	0,704	0,715	0,82	0,835	0,821	0,922	0,952
Halstead	0,742	0,652	0,801	0,767	0,5	0,932	0,588	0,738	0,826	0,848	0,804	0,718	0,958
McCabe	0,669	0,681	0,723	0,808	0,713	0,935	0,608	0,742	0,772	0,639	0,748	0,813	0,941
Misc	0,741	0,67	0,72	0,822	0,845	0,968	0,692	0,762	0,823	0,792	0,808	0,875	0,964

Table 15. Average AUC value on Random Forest

<i>Metrics</i>	<i>Average AUC</i>
LoC	0,808
Halstead	0,760
McCabe	0,753
Misc	0,806

3.6 K-Nearest Neighbor Classifier

Hasil klasifikasi tiap metrik pada setiap dataset menggunakan K-Nearest Neighbor dengan menggunakan $k = 3$ dapat dilihat dari Tabel 13. Dari rata-rata AUC pada Tabel 14 terlihat bahwa nilai klasifikasi terbaik pada K-Nearest Neighbor adalah metrik LOC.

Table 16. AUC Value on K-Nearest Neighbor

<i>Metrics</i>	<i>AUC Value</i>												
	<i>CMI</i>	<i>JMI</i>	<i>KCI</i>	<i>KC3</i>	<i>KC4</i>	<i>MC1</i>	<i>MC2</i>	<i>MW1</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>
LoC	0,683	0,5	0,731	0,665	0,672	0,754	0,656	0,696	0,714	0,695	0,703	0,795	0,878
Halstead	0,636	0,59	0,735	0,653	0,5	0,759	0,575	0,627	0,628	0,56	0,682	0,582	0,871
McCabe	0,528	0,451	0,701	0,686	0,722	0,752	0,638	0,704	0,62	0,519	0,628	0,659	0,87
Misc	0,624	0,546	0,688	0,727	0,805	0,78	0,631	0,664	0,728	0,51	0,723	0,787	0,9

Table 17. Average AUC value on K-Nearest Neighbor

<i>Metrics</i>	<i>Average AUC</i>
LoC	0,703
Halstead	0,646
McCabe	0,652
Misc	0,701

3.7 Support Vector Machine Classifier

Hasil klasifikasi setiap metrik pada setiap dataset dengan menggunakan Support Vector Machine dapat dilihat dari Tabel 15. Dari rata-rata AUC pada Tabel 16 terlihat bahwa nilai klasifikasi terbaik pada Support Vector Machine adalah metrik Misc.

Table 18. AUC Value on Support Vector Machine

<i>Metrik</i>	<i>Nilai AUC</i>												
	<i>CMI</i>	<i>JMI</i>	<i>KCI</i>	<i>KC3</i>	<i>KC4</i>	<i>MC1</i>	<i>MC2</i>	<i>MW1</i>	<i>PC1</i>	<i>PC2</i>	<i>PC3</i>	<i>PC4</i>	<i>PC5</i>
LoC	0,612	0,449	0,746	0,625	0,432	0,166	0,697	0,605	0,444	0,539	0,401	0,8	0,704
Halstead	0,466	0,558	0,734	0,518	0,65	0,183	0,676	0,562	0,604	0,58	0,475	0,521	0,728
McCabe	0,524	0,57	0,543	0,591	0,759	0,66	0,685	0,525	0,656	0,326	0,691	0,601	0,832
Misc	0,52	0,497	0,514	0,443	0,794	0,51	0,738	0,671	0,793	0,604	0,575	0,713	0,653

Table 19. Average AUC value on Support Vector Machine

<i>Metrics</i>	<i>Average AUC</i>
LoC	0,555
Halstead	0,558
McCabe	0,613
Misc	0,617

3.8 Discussion

The research result shows that each attribute on software metric categories has effect on each dataset. The comparison result of classifier performance based on software metric categories can be seen from Table 17 and Picture 2.

Table 20. Average AUC Software Metrics Value

<i>No</i>	<i>Classifier</i>	<i>Metrics</i>	<i>Average AUC</i>
1.	Naïve Bayes	LoC	0,756
		Halstead	0,746
		McCabe	0,743
		Misc	0,769
2.	Decision Tree	LoC	0,527
		Halstead	0,528
		McCabe	0,536
		Misc	0,566
3.	Random Forest	LoC	0,808
		Halstead	0,760
		McCabe	0,753
		Misc	0,806

4. K-Nearest Neighbor	LoC	0,703
	Halstead	0,646
	McCabe	0,652
	Misc	0,701
5. Support Vector Machine	LoC	0,555
	Halstead	0,558
	McCabe	0,613
	Misc	0,617

Picture 2. Average AUC Software Metrics Value Chart

Dari picture 2, Terlihat bahwa Algoritma Naive Bayes dan Algoritma Random Forest dapat memberikan performansi yang lebih baik dibandingkan algoritma lain dalam mengklasifikasikan perangkat lunak tertentu dengan pemilihan fitur berdasarkan metrik. Di sisi lain, kategori metrik terbaik pada setiap algoritme pengklasifikasi adalah metrik Misc.

5. CONCLUSION

Dari rata-rata nilai AUC dapat disimpulkan bahwa kategori metrik yang memberikan kinerja terbaik adalah metrik LoC, diikuti oleh metrik Misc. Kedua kategori tersebut telah mencapai nilai AUC tertinggi dalam pengklasifikasi Random Forest.

REFERENCES

- [1] R. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," J. Softw. Eng., vol. 1, 2015.
- [2] T. Menzies, J. Greenwald, and A. Frank, "Data Mining Static Code Attributes to Learn Defect Predictors.," IEEE Trans. Softw. Eng., vol. 33, pp. 2–13, 2007.
- [3] M. Kakkar and S. Jain, "Feature selection in software defect prediction: A comparative study," 2016, pp. 658–663.
- [4] T. M. Khoshgoftaar, K. Gao, and N. Seliya, "Attribute Selection and Imbalanced Data: Problems in Software Defect Prediction," in 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, 2010, vol. 1, pp. 137–144.

- [5] I. H. Laradji, M. Alshayeb, and L. Ghouti, "Software defect prediction using ensemble learning on selected features," *Inf. Softw. Technol.*, vol. 58, pp. 388–402, 2015.
- [6] Ö. F. Arar and K. Ayan, "A feature dependent Naive Bayes approach and its application to the software defect prediction problem," *Appl. Soft Comput.*, vol. 59, pp. 197–209, 2017.
- [7] E. Erturk and E. A. Sezer, "A comparison of some soft computing methods for software fault prediction," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1872–1879, 2015.