Available at https://jtsiskom.undip.ac.id (20 April 2021)
DOI:10.14710/jtsiskom.2021.13748

Jurnal Teknologi dan Sistem Komputer, 9(2), 2021, 77-82

Optimasi MWMOTE pada data tidak seimbang menggunakan complete linkage

MWMOTE optimization for imbalanced data using complete linkage

Meida Cahyo Untoro

Program Studi Teknik Informatika, Institut Teknologi Sumatera
Jl. Ryacudu, Lampung Selatan, Indonesia 35365

Cara sitasi: M. C. Untoro, "Optimasi MWMOTE pada data tidak seimbang menggunakan complete linkage," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 2, pp. 77-82, 2021. doi: 10.14710/jtsiskom.2021.13748, [Online].

Abstract – Imbalanced data can result in classification errors, such as in WMMOTE, and can decrease its performance and accuracy. Clustering in MWMOTE can be optimized to improve synthetic data generation and improve MWMOTE performance. This study aims to optimize the MWMOTE algorithm's performance in the clustering process in making synthetic data with complete linkage (CL). The dataset used a variety of data ratios to handle imbalanced data. The decision tree was used to determine the performance of MWMOTE and CL-MWMOTE oversampling. CL-MWMOTE evaluation results provide better and optimal performance than MWMOTE and increase the precision, recall, f-measure, and accuracy of 0.53 %, 0.67 %, 0.66 %, and 0.67 %, respectively.

Keywords *– imbalaced data; clustering; complete linkage; optimization; oversampling*

Abstrak – Data yang tidak seimbang dapat menyebabkan kesalahan klasifikasi, menurunkan kinerja dan akurasi. Pengelompokan pada MWMOTE dapat dioptimalkan untuk meningkatkan kinerja pembangkitan data sintetis menjadi representatif serta meningkatkan kinerja MWMOTE. Penelitian ini bertujuan untuk mengoptimalkan kinerja algoritme MWMOTE pada proses klasterisasi dalam pembuatan data sintetik dengan complete linkage (CL). Dataset yang digunakan memiliki beragam rasio data dengan tujuan menangangani data yang tidak seimbang. Decision tree digunakan untuk mengetahui kinerja dari oversampling MWMOTE dan CL-MWMOTE. Hasil evaluasi CL-MWMOTE memberikan kinerja yang lebih baik dan optimal daripada MWMOTE serta meningkatkan presisi sebesar 0,53 %, sensitivitas 0,67 %, f-measure 0,66 %, dan akurasi 0,67 %.

Kata kunci *– ketidakseimbangan data; complete linkage; klasterisasi; optimasi; oversampling*

*) Penulis korespondensi (Meida Cahyo Untoro)
Email: cahyo.untoro@if.itera.ac.id

I. PENDAHULUAN

Dalam sistem komputasi, ketidakseimbangan data (*imbalanced* data) merupakan pendistribusian data yang tidak seimbang. Jumlah data positif (mayoritas) yang lebih banyak dibandingkan dengan jumlah data negatif (minoritas) merupakan kondisi data yang tidak seimbang. Ketidakseimbangan data menimbulkan kejadian *misclassification*, dimana *classifier* lebih condong pada data mayoritas. Data minoritas akan dianggap sebagai *noise* dan *outlier* serta dapat menurunkan kinerja dari *classifier*.

Kasus ketidakseimbangan data salah satunya dapat ditangani dengan metode *oversampling* dan *undersampling* [1]. Kedua metode tersebut merupakan solusi ketidakseimbang data berdasarkan pengolahan atau penanganan pada bagian data [2]. Metode *oversampling* digunakan untuk menangani ketidakseimbangan data dengan cara membuat data sintetik pada data minoritas [3]. Metode *undersampling* melakukan pengurangan jumlah data mayoritas sampai sampai data memiliki distribusi jumlah data yang sama [4].

Oversampling menangani ketidakseimbangan data dengan cara membuat data sintetik yang terbentuk berdasarkan selisih jumlah data mayoritas dikurangi dengan data minoritas. Pembuatan data sintetik data minoritas bertujuan menyeimbangkan rasio antar kedua kelas tersebut. Data sintetik dibuat berdasarkan data minoritas dengan prinsip *k-Nearest Neighbor* (kNN) untuk mencari keanggotaan terdekat pada data minoritas, seperti dalam Synthetic Minority Over-Sampling Technique (SMOTE) [5].

Untoro dan Buliali [6] menyatakan bahwa data yang tidak seimbang pada data kesehatan dapat diselesaikan dengan Majority Weighted Minority Oversampling Technique (MWMOTE) dan memperoleh akurasi 85,47 %. Wei dkk. [7] melakukan modifikasi pada MWMOTE dengan meningkatkan kebisingan-imunitas (*noise-imunity*) yang bertujuan menangani kebisingan baru dan *overfitting*. Penggunaan *classifier* yang tepat dapat mempengaruhi hasil dari metode *oversampling* ini. Pohon keputusan (DT), kNN, naïve bayes dan SVM merupakan metode klasifikasi yang sering digunakan untuk memprediksi hasil ketidakseimbangan data [3], [8]. Metode klasifikasi yang

Submitted: 11 May 2020; Revised: 7 December 2020; Accepted: 18 January 2021; Published: 30 April 2020

tepat dalam permasalahan ketidakseimbangan data dengan MWMOTE adalah DT [8].

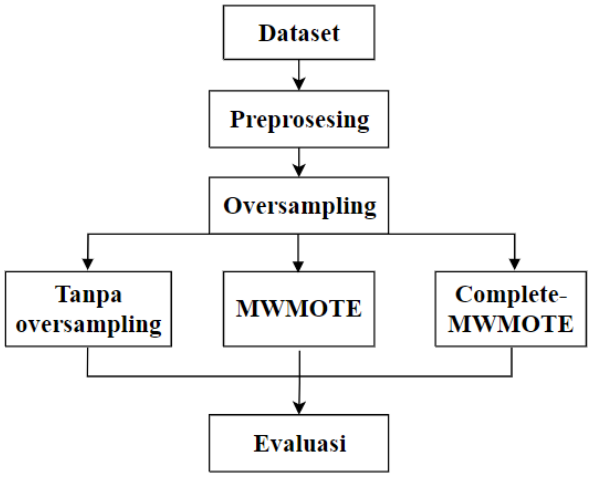
Lebih lanjut, Barua dkk. [9] mengusulkan MWMOTE sebagai metode perbaikan dari SMOTE melalui pembuatan data sintetik dengan pembobotan dan klasterisasi data minoritas. Pembobotan dan pengklasteran data minoritas ini bertujuan untuk mendapatkan data sintetik yang representatif. Hasil usulan tersebut mampu menurun derajat bias dan *overfitting* dan menghasilkan data sintetik dengan tingkat akurasi lebih baik dari proses klasterisasi. Namun, metode pengelompokan yang digunakan dalam MWMOTE merupakan bagian utama dalam menghasilkan data sintetis sehingga diperlukan metode pengelompokan yang lebih baik untuk mendapatkan performa yang terbaik.

Di sisi lain, *Agglomerative Hierarchy Clustering* (AHC) dapat menangani rasio ketidakseimbangan rendah pada data minoritas dan terjadi tumpang tindih data mayoritas [10]. Fikri dan Ulinnuha [11] menjelaskan bahwa *complete linkage* (CL) menjadi metode klasterisasi yang lebih baik dibandingkan *average linkage* berdasarkan nilai simpangan baku 0,222. Pengelompokan wajah dengan AHC dapat meningkatkan pengenalan wajah berbasis *Content Based Image Retrieval* (CBIR) dengan uji validasi 0,904938 berdasarkan *Cophenetic Correlation Coefficien* pada *complete linkage* [12]. Dari hal tersebut, penelitian ini bertujuan untuk mengoptimalkan metode klasterisasi yang berada dalam proses MWMOTE menggunakan *complete linkage*, selanjutnya disebut CL-MWMOTE, untuk membuat data sintetik lebih representatif dan menghasilkan klasifikasi data yang optimal. Hasil klasifikasi pada *oversampling* akan diuji dengan 10 dataset dari repositori UCI.

II. METODE PENELITIAN

Proses penelitian yang dilakukan dalam 3 tahap utama (Gambar 1). Tahap pertama dilakukan pembagian data latih dan uji dengan perbandingan data latih lebih banyak dari data uji. Data sintetik terbentuk dari proses *oversampling* MWMOTE, *Complete*-MWMOTE pada tahap kedua. Data sintetik berasal dari data minoritas pada data latih. Kinerja *oversampling* dievaluasi dengan klasifikasi menggunakan pohon keputusan (J48) berdasarkan nilai akurasi, presisi, sensitivitas, dan *f-measure*.

Metode klasterisasi pada penelitian menggunakan 10 dataset yang terdiri dari abalone, breast, ecoli, glass, libra, ocr, robot, satimage, wine, dan yeast (Tabel 1). Ratio data minoritas dan mayoritas pada 10 dataset berbeda antara satu data dengan dataset lainnya. Perbedaan rasio ketidakseimbangan data (IR) akan menentukan kinerja dari *complete linkage* MWMOTE (CL-MWMOTE) dalam membuat data sintetik pada semua dataset secara optimal. Rasio data mayoritas paling tinggi adalah dataset Abalone dengan 0,94:0,06 dan untuk data minoritas yang terbesar adalah dataset Breast dengan 0,34:0,66.



Gambar 1. Tahapan penelitian yang dilakukan

Tabel 1. Dataset UCI yang digunakan

Dataset	Atribut	Jumlah data	IR
Abalone	8	731	0,94 : 0,06
Breast	10	106	0,66 : 0,34
Ecoli	8	336	0,77 : 0,23
Glass	10	214	0,76 : 0,24
Libra	91	360	0,80 : 0,20
OCR	65	3823	0,90 : 0,10
Robot	25	5456	0,78 : 0,22
Satimage	37	6435	0,68 : 0,32
Wine	14	178	0,76 : 0,24
Yeast	9	1484	0,79 : 0,21

Tabel 2. Pemisahan data latih 70% dan data uji 30%.

Dataset	Pemisahan data		Jumlah data
	70%	30%	
Abalone	512	219	731
Breast	74	32	106
Ecoli	235	101	336
Glass	150	64	214
Libra	252	108	360
OCR	2676	1147	3823
Robot	3819	1637	5456
Satimage	4505	1931	6435
Wine	125	53	178
Yeast	1039	445	1484

Tahap pertama dalam penelitian adalah melakukan pembagian data menjadi data uji sebesar 30 % dan data latih 70 % (Tabel 2) [13]. Data latih digunakan sebagai pembuat model dari metode yang akan dievaluasi kinerjanya. Data uji digunakan untuk mengevaluasi hasil model yang telah dibuat oleh data latih .

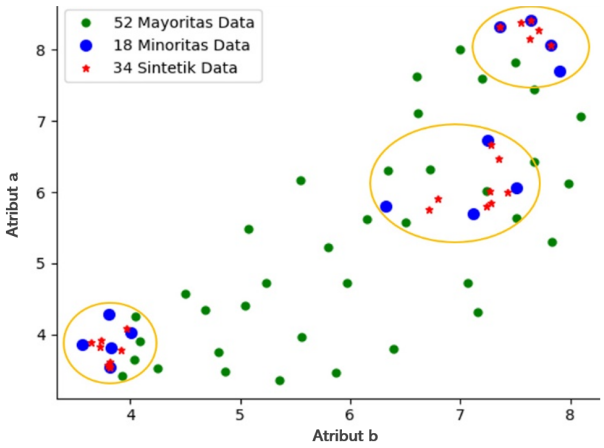
Tabel 3. Data minoritas dan mayoritas pada data latih

Dataset	Rasio data		Jumlah data
	Data mayoritas	Data minoritas	
Abalone	481	31	512
Breast	70	4	74
Ecoli	221	14	235
Glass	141	9	150
Libra	237	15	252
OCR	2516	161	2676
Robot	3590	229	3819
Satimage	4234	270	4505
Wine	117	7	125
Yeast	976	62	1039

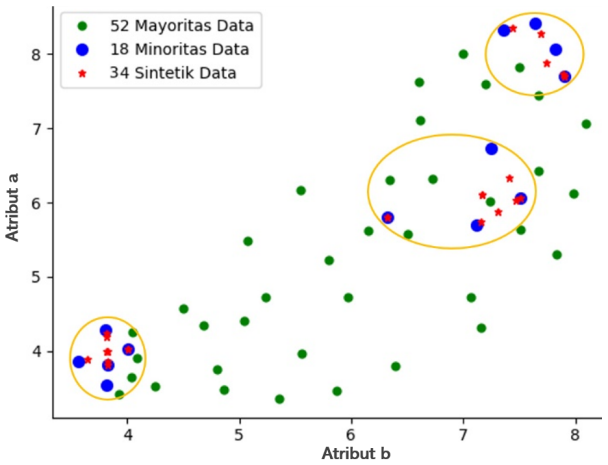
Tahap pembuatan data sintetik dilakukan dengan *oversampling* pada data minoritas sehingga data minoritas dan data mayoritas mempunyai jumlah data yang sama (data seimbang) (Tabel 3). Data sintetik dibuat berdasarkan data minoritas dengan tujuan menyeimbangkan data dengan cara pembobotan kelas minoritas dan melakukan klasterisasi untuk menentukan data minoritas yang akan dijadikan sebagai dasar pembuatan data sintetik. Data sintetik dibuat berdasarkan data minoritas dengan jumlah selisih dari data mayoritas dengan proses klasterisasi [7]. Pada MWMOTE, klasterisasi *average linkage* pada *oversampling* membentuk klaster berdasarkan jarak rata-rata antar klaster lainnya (Gambar 2) [14].

Di sisi lain, pembentukan dataset pada *complete linkage* bekerja berdasarkan jarak maksimum antar klaster lainnya (Gambar 3) [15]. Proses dalam CL-MWMOTE dinyatakan pada Algoritme 1 dan Gambar 4 Parameter D merupakan dataset yang tidak seimbang untuk dipisah menjadi data latih ($D_{m)}$ dan data uji (D_n) yang terdiri dari data mayoritas (D_{may}) dan data minoritas (D_{min}). D_{min} melakukan pencarian ketetanggaan terdekat berdasarkan jarak keanggotaan dengan algoritme kNN. Anggota dari D_{min} yang berada di antara D_{may} diberikan tanda *borderline* guna untuk memisahkan antar anggota dan menentukan D_{min} yang informatif. D_{min} yang informatif diberikan bobot berdasarkan probabilitas anggotanya. Pembobotan dilakukan untuk menjadikan kandidiat pembuatan data sintetik pada anggota D_{min}. Tahap terakhir dari metode CL-MWMOTE adalah pengklasteran. D_{min} yang sudah memiliki bobot dan mempunyai probabilitas akan dilakukan pembuatan data sintetik berdasarkan anggota D_{min}. Data sintetik dan D_{min} dengan jumlah rasio yang sama dengan D_{may}.

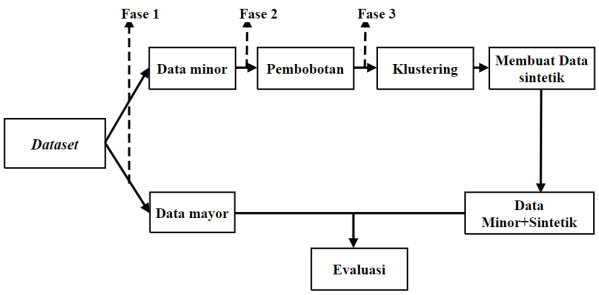
Proses *oversampling* dengan *complete linkage* pada CL-MWMOTE dan *average linkage* pada MWMOTE dievaluasi dengan pohon keputusan berdasarkan parameter spesifisitas, sensitivitas, akurasi, dan f-measure. Presisi menunjukkan ketepatan hasil prediksi berdasarkan metode yang diterapkan sesuai dengan dataset (Persamaan 1). Sensitivitas menunjukkan keberhasilan prediksi pada metode dalam menentukan kelas (Persamaan 2). Tingkat kedekatan antara nilai



Gambar 2. Pembentukan data sintetik dengan klasterisasi *average linkage* pada data sampel



Gambar 3. Pembentukan data sintetik dengan klasterisasi *complete linkage* pada data sampel



Gambar 4. Tahap algoritme CL-MWMOTE

prediksi dengan nilai nyata dinyatakan dengan akurasi (Persamaan 3). Nilai f-measure dinyatakan dengan Persamaan 4 yang mengkombinasikan presisi dan sensitivitas. *True positive* (TP) merupakan prediksi benar (data mayoritas) dan faktanya benar (data mayoritas). *True negative* (TN) memprediksi tidak benar (data minoritas) dan faktanya memang tidak benar (data minoritas). *False positive* (FP) memprediksi benar (data mayoritas), tetapi faktanya tidak benar (data minoritas). *False Negative* (FN) memprediksi tidak benar (data minoritas), tetapi faktanya adalah benar (data mayoritas).

Algoritme 1. CL-MWMOTE

Masukan:
D : Dataset
D_n : Dataset uji
D_m : Dataset latih
D_{min} : Data minoritas
D_{may} : Data mayoritas

1: **for** $x_i \in D_{min}$ **do**
//Memisahkan data minoritas dari data mayoritas
2: $D_{minf} = D_{min} - \{x_i \in D_{min} : DD(x_i)\}$
//Membuat batasan (borderline) antara data minoritas
// dengan mayoritas
3: $D_{bmay} = \cup_{x_i \in D_{minf}} D_{may}(x_i)$
//Memberikan label terhadap data minoritas yang
// informatif
4: $D_{min}(y_i) D D_{imin} = \cup_{y_i \in D_{bmay}} D_{min}(y_i)$
//Dataset yang terlabeli akan dilakukan pembobotan
//Pemilihan data minoritas berdasarkan probabilitas
5: $D_{mi}(x_i) = \sum_{y_i \in D_{my}} B_{mi}(y_i x_i)$
6: $Mi_p(x_i) = \frac{D_{mi}(x_i)}{\sum_{Z_i \in D_{min}} D_{mi}(Z_i)}$
end
//Pembuatan klaster berdasarkan D_{min}
7: D_{ps-min}
//Pembuatan data sintetik
8: **for** a = 1 **do**
9: $x \leftarrow D_{mi}$
10: $d = x + \propto \times (y - \propto) \rightarrow \propto = (0,1)$
11: $D_{omin} : D_{omin} = D_{omin} \cup \{d\}$
end
return d

$$Presisi = \frac{TP}{TP - FP} \quad (1)$$

$$Sensitivitas = \frac{TP}{TP + FN} \quad (2)$$

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F - measure = \frac{Presisi \times Sensitivitas}{Presisi + Sensitivitas} \quad (4)$$

III. HASIL DAN PEMBAHASAN

Hasil evaluasi kinerja klasifikasi terhadap data tak setimbang pada algoritme tanpa *oversampling* (normal), MWMOTE, dan CL-MWMOTE dengan pohon keputusan berdasarkan presisi dinyatakan pada Tabel 4, sensitivitas pada Tabel 5, *f-measure* pada Tabel 6, dan akurasi pada Tabel 7. Dataset Abalone, Breast, Glass, OCR, Satimage, dan Yeast mempunyai nilai presisi, sensitivitas, akurasi, dan f-*measure* yang lebih baik setelah dilakukan *oversampling* CL-MWMOTE.

Berdasarkan presisi, CL-MWMOTE mampu menangani ketidakseimbangan data berdasarkan enam dataset yang memberikan hasil lebih baik. MWMOTE mempunyai hasil presisi terbaik pada dataset Ecoli. Metode normal mempunyai presisi terbaik pada dataset Libra, Robot, dan Wine. Rerata presisi untuk ketiga metode ini adalah 95,86 %, 97,42 %, dan 97,94 %. CL-MWMOTE dapat meningkatkan kinerja presisi MWMOTE dengan rata-rata 0,53 % (Tabel 4).

Tabel 4. Kinerja presisi tanpa *oversampling*, MWMOTE, dan CL-MWMOTE

Dataset	Presisi (%)			
	Normal	**MWMOTE**	**CL-MWMOTE**	**Delta**
Abalone	96,80	98,42	**98,54**	0,12
Breast	83,60	91,66	**94,84**	3,18
Ecoli	92,00	**95,18**	94,56	-0,62
Glass	98,10	98,36	**98,94**	0,58
Libra	**99,20**	99,06	99,08	0,02
OCR	99,80	99,92	**99,96**	0,04
Robot	**99,90**	99,88	99,88	0
Satimage	98,50	99,90	**100**	0,1
Wine	**99,40**	98,26	99,28	1,02
Yeast	91,30	93,52	**94,36**	0,84
Rerata	*95,86*	*97,42*	*97,94*	*0,53*

Tabel 5. Kinerja sensitivitas tanpa *oversampling*, MWMOTE, dan CL-MWMOTE

Dataset	Sensitivitas (%)			
	Normal	**MWMOTE**	**CL-MWMOTE**	**Delta**
Abalone	96,90	97,78	**98,02**	0,24
Breast	82,10	88,64	**93,36**	4,72
Ecoli	92,00	**93,76**	92,84	-0,92
Glass	98,10	98,36	**98,92**	0,56
Libra	**99,18**	99,06	99,04	-0,02
OCR	99,80	99,92	**99,96**	0,04
Robot	**99,90**	99,88	99,88	0
Satimage	98,50	99,90	**100**	0,1
Wine	**99,40**	98,10	99,28	1,18
Yeast	91,60	93,00	**93,72**	0,84
Rerata	*95,75*	*96,84*	*97,5*	*0,67*

Seperti presisi, CL-MWMOTE juga mempunyai sensitivitas terbaik di enam dataset yang sama, sedangkan MWMOTE di dataset Ecoli dan normal di tiga dataset lainnya. Rerata sensitivitas untuk ketiga metode ini adalah 95,75 %, 96,84 %, dan 97,50 %. CL-MWMOTE dapat meningkatkan kinerja sensitivitas sebesar rata-rata 0,67 % (Tabel 5).

Nilai *f-measure* diperoleh melalui perbandingan hasil evaluasi antara parameter presisi dan sensitivitas. Hasil *oversampling* Ecoli menggunakan MWMOTE memiliki hasil *f-measure* tertinggi sebesar 94,02 % (Tabel 6), sedangkan dengan CL-MWMOTE sebesar 93,16 % dan dengan nilai normal (tanpa *oversampling*) sebesar 91,50 %. Hal tersebut disebabkan karena dataset Ecoli memiliki kemiripan karakteristik variabel atau atribut dan jarak persebaran data yang tidak terlalu jauh, atau dapat dikatakan homogen, sehingga tidak terlalu mempengaruhi hasil klasifikasi seperti dinyatakan [15]. Namun, secara umum CL-MWMOTE memiliki rerata *f-*

Tabel 6. Kinerja *f-measure* tanpa *oversampling*, MWMOTE, dan CL-MWMOTE

Dataset	F-measure (%)			
	Normal	**MWMOTE**	**CL-MWMOTE**	**Delta**
Abalone	96,40	97,96	**98,16**	0,2
Breast	80,60	88,98	**93,48**	4,5
Ecoli	91,50	**94,02**	93,16	-0,86
Glass	98,10	98,36	**98,94**	0,58
Libra	**99,20**	99,06	99,06	0
OCR	99,80	99,92	**99,96**	0,04
Robot	**99,90**	99,88	99,88	0
Satimage	98,50	99,90	**100**	0,1
Wine	**99,40**	98,10	99,28	1,18
Yeast	91,40	93,18	**93,94**	0,84
Rerata	*95,48*	*96,94*	*97,59*	*0,66*

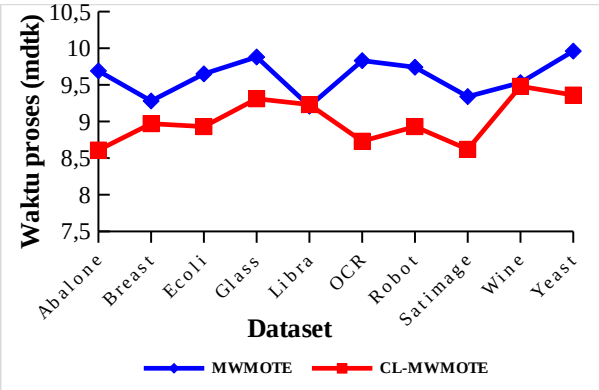
Tabel 7. Kinerja akurasi tanpa *oversampling*, MWMOTE, dan CL-MWMOTE

Dataset	Akurasi (%)			
	Normal	**MWMOTE**	**CL-MWMOTE**	**Delta**
Abalone	96,85	97,79	**98,03**	0,24
Breast	82,08	88,64	**93,33**	4,69
Ecoli	91,96	**93,75**	92,81	-0,94
Glass	98,13	98,36	**98,94**	0,58
Libra	**99,17**	99,05	99,05	0
OCR	99,79	99,92	**99,95**	0,03
Robot	99,85	99,88	99,88	0
Satimage	98,46	99,89	**100**	0,11
Wine	**99,44**	98,06	99,25	1,19
Yeast	91,58	93	**93,73**	0,84
Rerata	*95,73*	*96,83*	*97,5*	*0,67*

measure tertinggi sebesar 97,59 % dibandingkan tanpa *oversampling* 95,48 % dan MWMOTE sebesar 96,94 %.

Berdasarkan rerata akurasi, CL-MWMOTE memiliki kinerja yang lebih baik dibandingkan MWMOTE dan tanpa *oversampling*, yaitu sebesar 97,5 % dibandingkan 96,83 % dan 95,73 % (Tabel 7). Hasil rerata tersebut juga menunjukkan bahwa oversampling pada MWMOTE dan CL-MWMOTE dapat menangani ketidakseimbangan data dan meningkatkan kinerja akurasi sesuai [16], [17], selain kinerja presisi, sensitivitas, dan *f-measure*. Data sintetik yang dihasilkan oleh MWMOTE dengan basis klasterisasi *average linkage* memperlihatkan bahwa data sintetik yang dihasilkan dapat dioptimalkan dan lebih representatif [7]. Secara umum, CL-MWMOTE dapat menangani ketidakseimbangan data secara lebih optimal dibandingkan MWMOTE sesuai dengan [11], [12], dengan peningkatan akurasi rerata sebasar 0, 67 %.

Dari sisi waktu pemrosesan, CL-MWMOTE dapat menyelesaikan pemrosesan pada 10 dataset dalam waktu rata-rata 8,61 mdtk. Hasil performa MWMOTE terbaiknya adalah 9,34 mdtk dan terburuk 9,96 mdtk (Gambar 5). Hasil tersebut menunjukkan bahwa CL-MWMOTE mempunyai waktu proses yang lebih cepat daripada MWMOTE dalam menyeimbangkan data untuk membentuk data sintetik dan meningkatkan kinerja klasifikasinya. Peningkatan hasil kinerja CL-MWMOTE menunjukkan bahwa data sintetik yang terbentuk lebih representatif dibandingkan dengan MWMOTE dan tanpa *oversampling*.



Gambar 5. Hasil evaluasi kinerja waktu pada tanpa *oversampling*, MWMOTE dan CL-MWMOTE

IV. KESIMPULAN

Metode CL-MWMOTE memiliki kinerja yang lebih baik dibandingkan MWMOTE. Hasil evaluasi pohon keputusan CL-MWMOTE menunjukkan presisi 97,94 %, sensitivitas 97,50 %, akurasi 97,50 %, dan f-*measure* 97,59 %. Selain itu, waktu pemrosesan juga menunjukkan peningkatan dengan waktu proses 8,61 mdtk untuk 10 dataset. Hal tersebut menunjukkan bahwa klasterisasi *complete linkage* efektif meningkatkan hasil kinerja MWMOTE dan optimal dalam menangani ketidakseimbangan data.

DAFTAR PUSTAKA

[1] M. S. Shelke, P. R. Deshmukh, and P. V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *International Journal of Recent Trends in Engineering & Research*, vol. 3, no. 4, pp. 444–449, 2017. doi:10.23883/IJRTER.2017.3168.0UWXM

[2] T. Fahrudin, J. L. Buliali, and C. Fatichah, "Randshuff : an algorithm to handle imbalance class for qualitative data," *International Review on Computers and Software*, vol. 11, no.12, pp. 1093–1104, 2016. doi: 10.15866/irecos.v11i12.10956

[3] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE untuk menangani ketidakseimbangankelas dalam klasifikasi penyakit diabetes dengan C4.5, SVM, dan naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89-93, 2020. doi: 10.14710/jtsiskom.8.2.2020.89-93

[4] W. Y. Ng. Wing, J. Hu, D. S. Yeung, S. Yin, and F. Roli, "Diversified sensitivity-based undersampling for imbalance classification problems," *IEEE*

Transactions on Cybernetics, vol. 45, no. 11, pp. 2402–2412, 2015. doi: 10.1109/TCYB.2014.2372060

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953

[6] M. C. Untoro and J. L. Buliali, "Penanganan imbalance class data laboratorium kesehatan dengan Majority Weighted Minority Oversampling Technique," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 4, no. 1, p. 23, 2018. doi: 10.26594/register.v4i1.1184

[7] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "NI-MWMOTE : an improving noise-immunity majority weighted minority oversampling," *Expert Systems with Applications*, vol. 158, 113504, 2020. doi: 10.1016/j.eswa.2020.113504

[8] M. C. Untoro, M. Praseptiawan, and M. Widianingsih, "Evaluation of decision tree , k-nn , naive bayes and svm with mwmote on uci dataset evaluation of decision tree, k-nn, naive bayes and svm with mwmote on uci dataset," J*ournal of Physics: Conference Series*, vol. 1477, pp. 1–9, 2020. doi: 10.1088/1742-6596/1477/3/032005

[9] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority weighted minority oversampling technique for imbalanced dataset learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2014. doi: 10.1109/TKDE.2012.232

[10] C. Beyan and R. Fisher, "Classifying imbalanced datasets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653–1672, 2015. doi: 10.1016/j.patcog.2014.10.032

[11] S. Fikri and N. Ulinnuha, "Perbandingan metode single linkage, complete linkage dan average linkage dalam pengelompokan kecamatan berdasarkan variabel jenis ternak Kabupaten Sidoarjo," *Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 4, no. 2, pp. 1-5, 2019

[12] M. Fachrurrozi et al., "The grouping of facial images using agglomerative hierarchical clustering to improve the CBIR based face recognition system," in *International Conference on Data and Software Engineering*, Palembang, Indonesia, Nov. 2017, pp. 1–6. doi: 10.1109/ICODSE.2017.8285868

[13] S. Sivaranjani, S. Sivakumari, and S. Maragatham "GIS based serial crime analysis using data mining techniques," *International Journal of Computer Applications*, vol. 153, no. 8, pp. 19–23, 2016. doi: 10.5120/ijca2016912119

[14] Q. Nafisah and N. E. Chandra, "Analisis cluster average linkage berdasarkan faktor-faktor kemiskinan di Provinsi Jawa Timur," *Zeta - Math Journal*, vol. 3, no. 2, pp. 31–36, 2017. doi: 10.31102/zeta.2017.3.2.31-36

[15] M. Y. Pusadan, J. L. Buliali, and R. V. H Ginardi, "Anomaly detection of flight routes through optimal waypoint," in *International Conference Computer Application Informatics*, Medan, Indonesia, Dec. 2016, pp. 3–10. doi: 10.1088/1742-6596/801/1/012041

[16] W. Han, Z. Huang, S. Li, and Y. Jia, "Distribution-sensitive unbalanced data oversampling method for medical diagnosis," *Journal of Medical Systems*, vol. 43, 39, pp. 1–10, 2019. doi: 10.1007/s10916-018-1154-8

[17] R. Lotfian and C. Busso, "Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals," *IEEE Transactions on Affective Computing*, vol. XX, no. XX, 2019. doi: 10.1109/TAFFC.2019.2901465