# HYBRID APPROACH DENGAN EVOLUTIONARY HYBRID SAMPLING UNTUK PERMASALAHAN CLASS IMBALANCE DAN OVERLAPPING

*Hybrid Approach with Evolutionary Hybrid Sampling in Handling Class Imbalance and Overlapping*

**Teddy Surya Gunawan, Hartono, Sofyan Rahmad, Nurafni Damanik**
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
Universitas Potensi Utama
Program Studi Magister Ilmu Komputer
e-mail: hartonoibbi@gmail.com

***Abstrak***

*Permasalahan Class Imbalance merupakan permasalahan yang perlu mendapat penanganan serius di dalam proses klasifikasi. Permasalahan ini tidak dapat dihindari karena kecenderungan distribusi instance yang tidak seimbang yang mengakitbatkan suatu class memiliki instance yang jauh lebih besar dibandingkan class lainnya. Hal ini dapat mempengaruhi akurasi klasifikasi karena class dengan jumlah instance yang lebih besar memiliki akurasi yang lebih baik dibandingkan dengan class dengan jumlah instance yang lebih kecil. Penanganan class imbalance menggunakan pendekatan data-level, algorithm-level, dan hybrid approach. Hybrid approach yang menggabungkan data-level dan algorithm-level cenderung memberikan hasil yang lebih baik di dalam penanganan class imbalance. Di dalam Hybrid Approach penggunaan over-sampling dapat mengakitkan kondisi overlapping dan meaningless samples sedangkan under-sampling mengakibatkan hilangnya informasi penting dari majority samples. Pendekatan Evolutionary Hybrid Sampling digunakan untuk untuk menganalisa distribusi data pada original data dan memberikan area overlapping diantara majority dan minority dengan menggabungkan pendekatan over-sampling pada minority samples dan under-sampling pada majority samples. Penerapan Hybrid Approach dengan Evolutionary Hybrid Sampling diharapkan akan memberikan hasil yang lebih baik dibandingkan dengan Hybrid Approach dengan SMOTE sebagai metode Sampling. Hasil penelitian menunjukkan bahwa Hybrid Approach dengan Evolutionary Hybrid Sampling memberikan hasil yang lebih baik pada Augmented R-Value, Precision, dan Recall.*

***Kata kunci***—Class Imbalance, Hybrid Approach, Over-Sampling, Under-Sampling, Evolutionary-Hybrid Sampling

***Abstract***

*Class Balance problems need serious handling in the classification process. This problem cannot be avoided because of the tendency of an unequal distribution of instances which is indicated by the existence of a class having much larger instances than other classes. This can affect classification accuracy because classes with a larger number of instances tend to have better accuracy than classes with a smaller number of instances. The method of handling class imbalance can be done using data-level, algorithm-level, and hybrid approaches. Hybrid approach that combines data-level and algorithm-level tends to give better results in handling class imbalance. In the Hybrid Approach the use of over-sampling can result in overlapping conditions and meaningless samples while under-sampling results in the loss of important information from the majority samples. Evolutionary Hybrid Sampling approach is used to analyze the distribution of data on the original data and provide overlapping areas between majority and minority by combining over-sampling on minority samples and under-sampling on majority samples. Hybrid Approach with Evolutionary Hybrid Sampling is expected to give better results than the Hybrid*

*Approach with SMOTE as the sampling method. The results showed that the Hybrid Approach with Evolutionary Hybrid Sampling gave better results on Augmented R-Value, Precision, and Recall.*

***Keywords***— Class Imbalance, Hybrid Approach, Over-Sampling, Under-Sampling, Evolutionary-Hybrid Sampling

## 1. PENDAHULUAN

Permasalahan *class imbalance* merupakan permasalahan yang umum dihadapi di dalam proses klasifikasi disebabkan oleh *dataset* di dalam dunia nyata yang pada umumnya memiliki distribusi data yang tidak seimbang. Permasalahan ini ditandai dengan terdapat suatu *class* dengan jumlah *instance* yang jauh lebih besar (*majority class*) dibandingkan dengan *class* lainnya(*minority class*) (Ahsan et al., 2022). Adanya permasalahan *class imbalance* ini menyebabkan akurasi yang lebih rendah pada *class* pada *minority class* dan juga menyebabkan tidak diperolehnya informasi penting pada *minority class* yang seringkali merupakan class yang mengandung informasi yang menarik(Shin et al., 2020).

Terdapat 3 (tiga) kategori pendekatan yang umum digunakan di dalam penanganan *class imbalance,* yakni: *data-level, algorithm-level,* dan *hybrid approach(De Angeli et al., 2022)*. *Data-level* dilakukan dengan mengubah distribusi data pada tiap *class* dengan cara menaikkan jumlah *samples* pada *minority class* (*over-sampling*) dan menurunkan jumlah *samples* pada *majority class* (*under-sampling*)(Wang & Cheng, 2021). Adapun *algorithm-level* berperan dengan membangkitkan sejumlah *classifier* yang akan membantu di dalam proses *rebalancing samples* pada tiap *class*(Mienye & Sun, 2021). *Hybrid Approach* merupakan pendekatan yang menggabungkan *data-level* dan *algorithm-level* dan memiliki kecenderungan untuk memberikan hasil yang lebih baik di dalam penanganan *class imbalance*(Czarnowski, 2022).

Pemilihan metode *sampling* yang tepat merupakan kunci keberhasilan pada *Hybrid Approach* dengan kecenderungan bahwa lebih banyak peneliti yang tertarik menggunakan metode *over-sampling* khususnya Synthetic Minority Over-Sampling Technique (SMOTE)(Soltanzadeh & Hashemzadeh, 2021). Permasalahan utama pada penggunaan *over-sampling* adalah bahwa proses *oversampling* pada *minority class* dapat menghasilkan *meaningless samples* dan juga *overlapping*(de Morais & Vasconcelos, 2019). Adapun *under-sampling* pada *majority class* sering menyebabkan hilangnya informasi penting pada *majority class*(Bach et al., 2019).

Penggabungan metode *over-sampling* dan *under-sampling* telah mendapat perhatian dari sejumlah peneliti. Penelitian yang dilakukan oleh (Hanskunatai, 2018) telah menggabungkan penggunaan *over-sampling* dengan SMOTE dan juga *Random Under Sampling* dan memperoleh hasil bahwa terdapatnya kesulitan untuk memperoleh *overlapping area* antara *majority class* dan *minority class*. Penelitian yang dilakukan oleh (Xu et al., 2020) menggabungkan penerapan *M-SMOTE* sebagai metode *over-sampling* pada *minority class,* yang digabungkan dengan metode *Edited Nearest Neighbor* (ENN) untuk menghilangkan data yang tidak penting (*noise*) pada *majority class*. Kesulitan yang dihadapi metode ini adalah berupa penentuan *stopping criteria* yang berkaitan dengan terjadinya kondisi *local optima*.

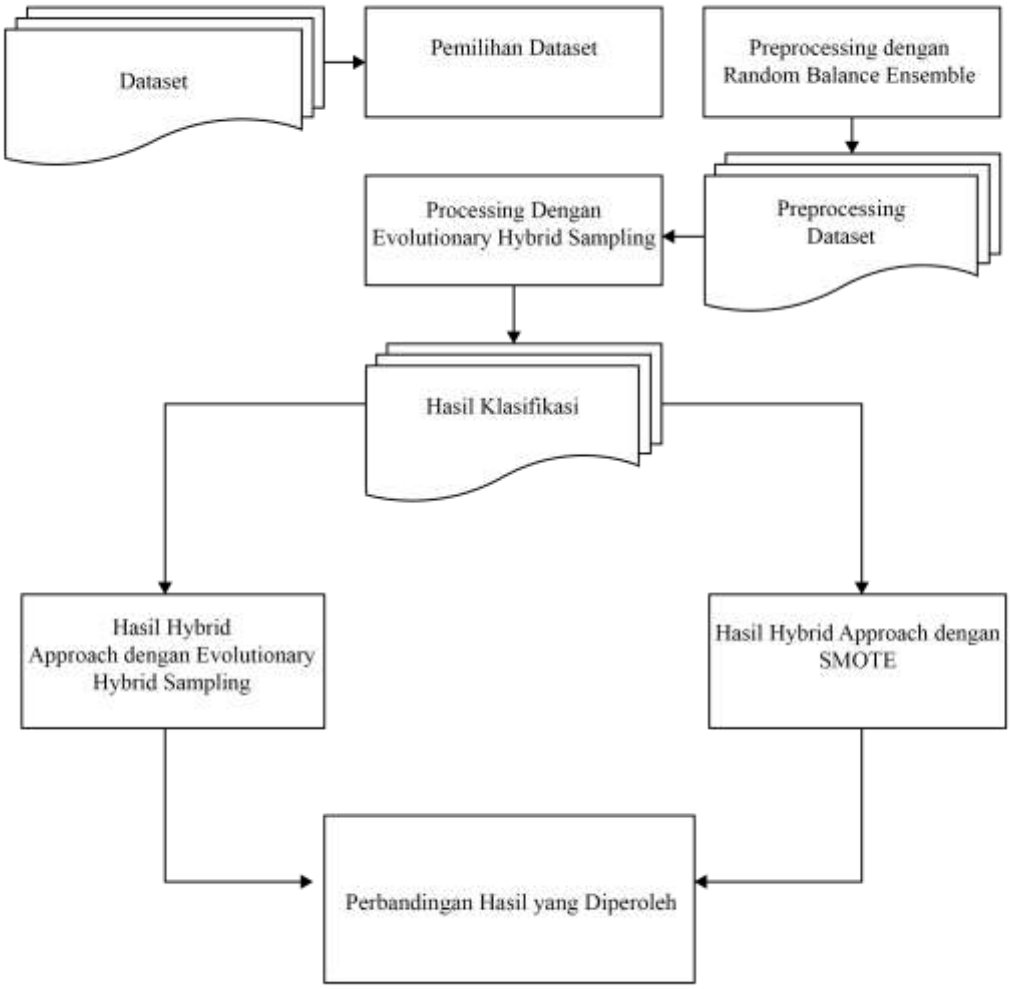
Metode *Evolutionary Hybrid Sampling* yang dikemukakan oleh (Zhu et al., 2020) yang menggabungkan penggunaan *CHC Algorithm* dengan kombinasi *over-sampling* dan *under-sampling* dapat dipertimbangkan sebagai metode *sampling* di dalam penangagan *class imbalance*. CHC *algorithm* merupakan varian dari *genetic algorithm* yang pertama kali dikemukakan oleh (Eshelman, 1991). Metode ini menawarkan keunggulan berupa kemampuan untuk menentukan *overlapping region* dan juga penentuan *stopping criteria.*

Penelitian ini akan menguji hasil yang diperoleh oleh *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* untuk Permasalahan *Class Imbalance* dan *Overlapping*.

## 2. METODE PENELITIAN

### *2.1 Tahapan Penelitian*

Adapun tahapan dari penelitian dapat dilihat pada Gambar 1.




Gambar 1. Tahapan Penelitian

Berdasarkan pada Gambar 1 dapat dilihat bahwa pelaksanaan penelitian dimulai dengan penentuan *dataset* yang akan digunakan. *Dataset* yang akan digunakan adalah *dataset* yang mengandung permasalahan *class imbalance* yang diperoleh dari *KEEL Repository*(Alcalá-Fdez et al., 2009). Tahapan pertama dari penanganan *class imbalance* adalah *preprocessing* yang akan dilakukan dengan menggunakan *Random Balance Ensemble*. Metode *Random Balance Ensemble* dilakukan dengan cara membangkitkan sejumlah *classifier* yang kemudian akan secara acak akan melakukan proses *over-sampling* pada *minority class* dan juga secara acak akan melakukan proses *under-sampling* pada *majority class*.

*Processing* dilakukan dengan menggunakan *Evolutionary Hybrid Sampling*, dimana *Evolutionary Hybrid Sampling* dilakukan untuk menentukan *Global Optimal Solution* yang

menjadi dasar dari proses *under-sampling* pada *majority class* dan juga sekaligus mencegah terjadinya *overlapping* pada *majority* dan *minority class*. Hasil penelitian ini akan dibandingkan dengan *Hybrid Approach* dengan SMOTE.

### *2.2 Hybrid Approach*

Adapun *pseudocode* dari *Hybrid Approach* adalah sebagai berikut(Galar et al., 2012).

$Input: D_T = \{x_1, x_2, \dots, x_n\} // Training\ Dataset$
$N = Number\ of\ Classifier$
$Output: Classification\ Prediction\ P$
$Method:$
$Step\ 1\ Preprocessing\ using\ Preprocessing\ Method$
$Step\ 2\ For\ i = 1\ to\ N\ do$
$\quad i. Apply\ Machine\ Learning\ Classification\ Algorithm\ on\ The\ Attributes\ of\ D_T$
$\quad ii. Obtain\ Classification\ Prediction\ P_i\ from\ machine\ learning\ classification\ algorithm$
$End\ For$
$Step\ 3\ For\ i = 1\ to\ n$
$\quad Apply\ processing\ using\ bagging, boosting\ or\ sampling$
$End\ For$

Berdasarkan pada *pseudocode* dapat dilihat bahwa pada *Hybrid Approach* secara umum tahapan yang dilalui ada 2 (dua), yaitu: tahapan *preprocessing* dan tahapan *preprocessing*. Baik pada tahapan *preprocessing* maupun tahapan *processing* dilakukan dengan metode yang menggunakan baik pendekatan *data-level* yang meliputi proses *sampling* maupun pendekatan *algorithm-level* yang membangkitkan sejumlah *classifier*.

### *2.3 Evolutionary Hybrid Sampling*

Adapun *pseudocode* dari *Evolutionary Hybrid Sampling* adalah sebagai berikut.

$Input: Dataset\ S, Population\ M, Maximum\ Iteration\ Times\ T, Nearest\ Neighbors\ K$
$Output: Rebalanced\ Dataset\ S'$
$Get\ the\ Set\ of\ Majority\ Class\ Maj\ in\ S$
$For\ x_i\ in\ Maj\ do$
$\quad Get\ its\ k\ nearest\ neighbors\ x_i^k$
$\quad for\ x_{i\text{-}neighbors}\ in\ x_i^k\ do$
$\quad\quad if\ x_{i\text{-}neighbors}\ \varepsilon \min then$
$\quad\quad\quad Add\ x_i\ to\ overlapping\ set\ o\_set$
$\quad\quad\quad break$
$\quad\quad end$
$\quad end$
$end$
$For\ i = 1\ to\ M\ do$
$\quad Initialize\ Chromosome\ c_i\ for\ samples\ in\ o_{set} and\ join\ in\ population\ P_{t\text{-}1}$
$\quad Calculate\ The\ Fitness\ Value\ of\ Chromosome\ c_i$
$end$
$Select\ The\ Chromosome\ with\ Maximum\ Fitness\ as\ global\ optimal\ solution\ R_{t\text{-}1}$
$Break_{Flag} = 0$
$While\ Break_{Flag} < 10\ do$
$\quad For\ t = 1\ to\ T\ do$
$\quad\quad For\ \forall c_i, c_j \epsilon\ P_{t\text{-}1}\ do$
$\quad\quad\quad If\ Hamming\ Distance\ of\ c_i \text{ and } c_j > Threshold\_HUX\ then$
$\quad\quad\quad\quad Generate\ Two\ Child\ Chromosome\ and\ Join\ to\ Contemporary\ Population\ P_t$

*End*
*End*
*If* $P_t$ *is empty then*
$Threshold_{HUX} = Threshold_{HUX}$-1
*End*
*Calculate Fitness Value Using Equation* 1 *until* 4

$$IR = \frac{\Delta_{maj}}{\Delta_{min}} \quad (1)$$

$$OR = \sum_{i=1}^{maj} \Delta_{min}^{xi,k}/K.\Delta_{maj}^{O}, x_i \in Majority \quad (2)$$

$$GM = \sqrt{TP_{rate}.TN_{rate}} \quad (3)$$

$$Fitness = \propto.\frac{1-OR}{IR}+(1-\propto).\mathrm{GM} \quad (4)$$

*Select the first M chromosome from* $P_{t-1} \cup P_t$ *with maximum fitness value as* $P_t$
*Select The Chromosome with Maximum Fitness as Global Optimal Solution* $R_t$
*If* $Global\_Fitness_t == Global\_Fitness_{t-1}$ *then*
*Mutate with Mutation Ratio*
$Break\_Flag = Break\_Flag + 1$
*End*
*Else*
$Break\_Flag = 0$
*End*
*End*
*End*
*Undersample Majority Samples with Global Optimal Solution* $R_t$
*Randomly Oversample all the minority samples in S*
*Return rebalanced dataset* $S'$

Berdasarkan pada *pseudocode* dapat dilihat bahwa proses *evolutionary hybrid sampling* dimulai dengan penanganan *overlapping* yang dilakukan dengan mengecek apakah terdapat area pada *majority class* yang *overlapping* dengan *minority class*, bila terdapat *samples* atau *instance* yang *overlapping* maka akan dimasukkan ke dalam himpunan *overlapping* $o\_set$. Apabila sudah tidak terdapat area yang *overlapping* maka proses akan berlanjut ke *evolutionary algorithm* yang dimulai dengan pembangkitan kromosom. Kemudian akan dihitung nilai *fitness value* dari masing-masing kromosom. Kromosom dengan nilai *fitness* tertinggi akan menjadi *global optimal solution*. Kemudian lakukan proses *under-sapling* pada *majority class* dengan nilai *global optimal solution*. Hal ini akan dilanjutkan dengan proses *over-sampling* terhadap seluruh *minority samples* sehingga akan diperoleh *dataset* yang telah seimbang.

### *2.4 Parameter Pengukuran*

#### *2.4.1 Confusion Matrix*

*Confusion Matrix* adalah matriks yang menggambarkan pengukuran hasil klasifikasi untuk tiap *positive samples* maupun *negative samples.* Adapun *Confusion Matrix* dapat dilihat pada Tabel 1(Luque et al., 2019).

Tabel 1. *Confusion Matrix*

| | **Predictive Positive Class** | **Predictive Negative Class** |
|---|---|---|
| Actual Positive Class | True Positive (TP) | False Negative (FN) |
| Actual Negative Class | False Positive (FP) | True Negative (TN) |

#### *2.4.2 Augmented R-Value*

*Augmented R-Value* menggambarkan seberapa besar *overlapping* terjadi. Semakin besar *Augmented R-Value* maka semakin besar *overlapping* terjadi. *Augmented R-Value* dapat diukur dengan menggunakan Persamaan 5(Oh, 2011).

$$R_{aug}(D[V]) = \frac{\sum_{i=0}^{k-1} |C_{k-1-i}| R(C_i)}{\sum_{i=0}^{k-1} |C_i|} \quad (5)$$

Dimana $C_0, C_1, \ldots, C_{k-1}$ adalah *k* class labels dengan $|C_0| \geq |C_1| \geq \cdots \geq |C_{k-1}|$ dan $D[V]$: Dataset D yang mengandung *predictor* di dalam himpunan *V*. Semakin besar $R_{Aug}$ maka semakin besar *overlap* yang terjadi.

*2.4.3 Precision dan Recall*

*Precision* menyatakan tingkat akurasi dari hasil prediksi terhadap *positive samples* yang berhasil diberikan oleh hasil klasifikasi dengan membandingkan hasil yang diperoleh dengan kesalahan klasifikasi dari *negative samples* menjadi *positive samples*. Adapun *Recall* menyatakan tingkat akurasi hasil prediksi *positive samples* dengan membandingkan hasil yang diperoleh dari kesalahan klasifikasi *positive samples* menjadi *negative samples.* Adapun *Precision* dan *Recall* dapat ditentukan dengan menggunakan Persamaan 6 dan 7.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

*2.5 Dataset yang Digunakan*

*Dataset* yang digunakan di dalam penelitian ini bersumber dari *KEEL Repository*(Alcalá-Fdez et al., 2009). Adapun *dataset* yang digunakan dapat dilihat pada Tabel 2.

Tabel 2. *Dataset* yang Digunakan

| *Dataset* | Jumlah Atribut | Jumlah *Instance* | *Imbalance Ratio* |
|---|---|---|---|
| Glass1 | 9 | 214 | 1.82 |
| Wisconsin | 9 | 683 | 1.86 |
| Glass 0 | 9 | 214 | 2.06 |
| Ecoli2 | 7 | 336 | 5.46 |
| Yeast3 | 8 | 1484 | 8.1 |
| Page-blocks0 | 10 | 5472 | 8.79 |

Berdasarkan pada Tabel 2 dapat dilihat bahwa terdapat 6(enam) *dataset* dengan jumlah atribut, jumlah *instance*, dan juga *imbalance ratio* yang beragam. *Dataset* ini yang akan diuji di dalam penelitian ini dengan menggunakan *stratified K-Fold* (K=10).

## 3. HASIL DAN PEMBAHASAN

*3.1 Pengujian untuk Augmented R-Value*

Adapun hasil pengujian dengan *Augmented R-Value* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Pengujian untuk *Augmented R-Value*

| *Dataset* | *Augmented R-Value* untuk *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* | *Augmented R-Value* untuk *Hybrid Approach* dengan SMOTE |
|---|---|---|
| Glass1 | 0.271 | 0.272 |
| Wisconsin | 0.278 | 0.287 |
| Glass 0 | 0.287 | 0.292 |
| Ecoli2 | 0.291 | 0.302 |
| Yeast3 | 0.301 | 0.311 |

| | | |
|---|---|---|
| Page-blocks0 | 0.312 | 0.314 |

Berdasarkan pada Tabel 3 dapat dilihat bahwa hasil yang diberikan oleh *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* dan *Hybrid Approach* dengan SMOTE sudah cukup baik di dalam penanganan *overlapping* yang ditunjukkan dengan nilai *Augmented R-Value* yang relatif kecil. Secara umum dapat dikatakan bahwa hasil yang diberikan oleh *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* lebih baik jika dibandingkan dengan *Hybrid Approach* dengan SMOTE. Hasil penelitian menunjukkan bahwa tinggi nilai *imbalance ratio* maka kualitas hasil penanganan *overlapping* yang diberikan oleh kedua metode cenderung mengalami penurunan.

### *3.2. Pengujian untuk Precision*

Adapun hasil pengujian nilai *Precision* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Pengujian untuk *Precision*

| *Dataset* | *Precision* untuk *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* | *Precision* untuk *Hybrid Approach* dengan SMOTE |
|---|---|---|
| Glass1 | 0.95 | 0.94 |
| Wisconsin | 0.92 | 0.91 |
| Glass 0 | 0.88 | 0.87 |
| Ecoli2 | 0.86 | 0.81 |
| Yeast3 | 0.84 | 0.83 |
| Page-blocks0 | 0.81 | 0.79 |

Berdasarkan pada Tabel 4 dapat dikatakan bahwa nilai *Precision* yang diperoleh oleh kedua metode sudah baik. Berdasarkan pada Tabel 4 dapat dilihat bahwa *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* memberikan hasil yang lebih baik jika dibandingkan dengan *Hybrid Approach* dengan SMOTE. Hasil ini menunjukkan bahwa semakin tinggi nilai *imbalance ratio* maka hasil yang diperoleh akan mengalami penurunan.

### *3.3. Pengujian untuk Recall*

Adapun hasil pengujian nilai *Recall* dapat dilihat pada Tabel 5.

Tabel 5. Hasil Pengujian untuk *Recall*

| *Dataset* | *Precision* untuk *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* | *Precision* untuk *Hybrid Approach* dengan SMOTE |
|---|---|---|
| Glass1 | 0.91 | 0.92 |
| Wisconsin | 0.93 | 0.91 |
| Glass 0 | 0.87 | 0.81 |
| Ecoli2 | 0.88 | 0.79 |
| Yeast3 | 0.87 | 0.82 |
| Page-blocks0 | 0.77 | 0.75 |

Berdasarkan pada Tabel 5 dapat dikatakan secara umum *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* memberikan hasil yang lebih baik jika dibandingkan dengan *Hybrid Approach* dengan SMOTE. Pengecualian hanya pada *dataset* Glass1 dimana hasil yang diberikan oleh *Hybrid Approach* dengan SMOTE lebih baik. Pada kedua metode terdapat kecenderungan bahwa hasil yang diperoleh mengalami penurunan jika nilai *imbalance ratio* semakin meningkat.

### *3.4. Pengujian Tingkat Signifikansi*

Adapun pengujian tingkat signifansi dilakukan dengan menggunakan *Wilcoxon Signed-Rank Test*. Adapun hasil pengujian tingkat signifikansi perbedaan dapat dilihat pada Tabel 6.

Tabel 6. Tingkat Signifikansi Dengan Wilcoxon Signed-Rank Test.

| *Parameter* | *P-Value* | *Kesimpulan* |
|---|---|---|
| Augmented R-Value | 0.0312500 | Terdapat perbedaan signifikan karena nilai P<0.05 |
| Precision | 0.0310325 | Terdapat perbedaan signifikan karena nilai P<0.05 |
| Recall | 0.0584753 | Tidak Terdapat perbedaan signifikan karena nilai P>0.05 |

Berdasarkan pada Tabel 6 dapat dilihat bahwa perbedaan signifikan antara kedua metode hanya ditunjukkan pada nilai *Augmented R-Value* dan *Precision*. Adapun untuk *Recall* perbedaan yang ditunjukkan oleh kedua metode tidak signifikan.

## 4. KESIMPULAN

Adapun kesimpulan dari hasil penelitian ini adalah sebagai berikut.

1. Hasil penanganan *overlapping* yang diberikan oleh kedua metode sudah cukup baik yang ditunjukkan dengan nilai *Augmented R-Value* yang sudah cukup rendah yang menunjukkan bahwa *overlapping* yang terjadi juga tidak besar. Hasil penelitian menunjukkan bahwa hasil yang diberikan oleh *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* lebih baik jika dibandingkan dengan *Hybrid Approach* dengan SMOTE.
2. Nilai *precision* yang diberikan oleh *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* lebih baik jika dibandingkan dengan *Hybrid Approach* dengan SMOTE.
3. Nilai *Recall* yang diberikan oleh *Hybrid Approach* dengan *Evolutionary Hybrid Sampling* lebih baik jika dibandingkan dengan *Hybrid Approach* dengan SMOTE. Pengecualian hanya pada *dataset* Glass 1.
4. Pada kedua metode diperoleh bahwa semakin besar nilai *imbalance ratio* maka hasil yang diperoleh mengalami penurunan baik untuk nilai *Augmented R-Value, Precision*, maupun *Recall.*
5. Perbedaan signifikan untuk kedua metode hanya ditunjukkan pada nilai *Augmented R-Value* dan *Precision*. Adapun untuk nilai *Recall* perbedaan yang ada tidak signifikan.

## 5. SARAN

Penelitian mendatang diharapkan dapat dikembangkan untuk menangani permasalahan *multi-class imbalance* dan juga peningkatan kemampuan di dalam menangani *dataset* dengan *imbalance ratio* yang tinggi.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Rektor Universitas Potensi Utama dan Ketua Program Studi Magister Ilmu Komputer Universitas Potensi Utama atas dukungan yang diberikan hingga terselesaikannya penelitian ini.

## DAFTAR PUSTAKA

[1] Ahsan, R., Ebrahimi, F., & Ebrahimi, M. (2022). Classification of imbalanced protein sequences with deep-learning approaches; application on influenza A imbalanced virus classes. *Informatics in Medicine Unlocked*, 100860. https://doi.org/10.1016/j.imu.2022.100860

[2] Alcalá-Fdez, J., Sánchez, L., García, S., Jesus, M. J. del, Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., Fernández, J. C., & Herrera, F. (2009). KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, *13*(3), 307–318. https://doi.org/10.1007/s00500-008-0323-y

[3] Bach, M., Werner, A., & Palt, M. (2019). The Proposal of Undersampling Method for Learning from Imbalanced Datasets. *Procedia Computer Science*, *159*, 125–134. https://doi.org/10.1016/j.procs.2019.09.167

[4] Czarnowski, I. (2022). Weighted Ensemble with one-class Classification and Over-sampling and Instance selection (WECOI): An approach for learning from imbalanced data streams. *Journal of Computational Science*, *61*, 101614. https://doi.org/10.1016/j.jocs.2022.101614

[5] De Angeli, K., Gao, S., Danciu, I., Durbin, E. B., Wu, X.-C., Stroup, A., Doherty, J., Schwartz, S., Wiggins, C., Damesyn, M., Coyle, L., Penberthy, L., Tourassi, G. D., & Yoon, H.-J. (2022). Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types. *Journal of Biomedical Informatics*, *125*, 103957. https://doi.org/10.1016/j.jbi.2021.103957

[6] de Morais, R. F. A. B., & Vasconcelos, G. C. (2019). Boosting the performance of over-sampling algorithms through under-sampling the minority class. *Neurocomputing*, *343*, 3–18. https://doi.org/10.1016/j.neucom.2018.04.088

[7] Eshelman, L. J. (1991). The CHC Adaptive Search Algorithm: How to Have Safe Search When Engaging in Nontraditional Genetic Recombination. In G. J. E. Rawlins (Ed.), *Foundations of Genetic Algorithms* (Vol. 1, pp. 265–283). Elsevier. https://doi.org/10.1016/B978-0-08-050684-5.50020-3

[8] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

[9] Hanskunatai, A. (2018). A New Hybrid Sampling Approach for Classification of Imbalanced Datasets. *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, 67–71. https://doi.org/10.1109/CCOMS.2018.8463228

[10] Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/j.patcog.2019.02.023

[11] Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, *25*, 100690. https://doi.org/10.1016/j.imu.2021.100690

[12] Oh, S. (2011). A new dataset evaluation method based on category overlap. *Computers in Biology and Medicine*, *41*(2), 115–122. https://doi.org/10.1016/j.compbiomed.2010.12.006

[13] Shin, J., Yoon, S., Kim, Y., Kim, T., Go, B., & Cha, Y. (2020). Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics*, 101202. https://doi.org/10.1016/j.ecoinf.2020.101202

[14] Soltanzadeh, P., & Hashemzadeh, M. (2021). RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Information Sciences*, *542*, 92–111. https://doi.org/10.1016/j.ins.2020.07.014

[15] Wang, Y.-C., & Cheng, C.-H. (2021). A multiple combined method for rebalancing medical data with class imbalances. *Computers in Biology and Medicine*, *134*, 104527. https://doi.org/10.1016/j.compbiomed.2021.104527

[16] Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*, 103465. https://doi.org/10.1016/j.jbi.2020.103465

[17] Zhu, Y., Yan, Y., Zhang, Y., & Zhang, Y. (2020). EHSO: Evolutionary Hybrid Sampling in overlapping scenarios for imbalanced learning. *Neurocomputing*, *417*, 333–346. https://doi.org/10.1016/j.neucom.2020.08.060