

# KOMPARASI METODE SMOTE-TOMEK DAN SMOTE-ENN UNTUK MENGATASI DATA TIDAK SEIMBANG

Fandi Yulian Pamuji<sup>1</sup>, Andriyan Rizki Jatmiko<sup>2</sup>, Mohammad Dwi Irfan Affandi<sup>3</sup>

<sup>1,2,3</sup> Sistem Informasi, Fakultas Teknologi Informasi, Universitas Merdeka Malang, Indonesia  
<sup>1</sup>fandi.pamuji@unmer.ac.id, <sup>2</sup>andriyan.jatmiko@unmer.ac.id, <sup>3</sup>22083000105@student.unmer.ac.id

## Abstrak

Perkembangan teknologi informasi dan data analisis saat ini mendorong meningkatnya pemanfaatan *machine learning* dalam berbagai bidang. Namun, salah satu permasalahan umum yang sering muncul dalam penerapan *machine learning* adalah kondisi data tidak seimbang (*imbalanced data*) yaitu ketidakseimbangan jumlah data antar kelas, di mana kelas mayoritas jauh lebih dominan dibandingkan kelas minoritas. Untuk mengatasi permasalahan *dataset* tidak seimbang adalah dengan menyeimbangkan distribusi kelas tidak seragam di antara kelas-kelas dengan komparasi menggunakan metode SMOTE-Tomek dan SMOTE-ENN supaya jumlahnya seimbang dari kelas mayoritas (negatif) maupun kelas minoritas (positif). Berdasarkan hasil eksperimen yang telah dilakukan dari penelitian ini yaitu bahwa pengujian metode SMOTE-Tomek dengan metode klasifikasi mampu menangani jumlah kelas mayoritas (negatif) dan kelas minoritas (positif) pada data tidak seimbang dengan menghasilkan nilai MCC dan G-mean mencapai kinerja prediksi yang lebih besar dibandingkan dengan menggunakan metode klasifikasi saja maupun menggunakan Metode SMOTE-ENN. Kemudian untuk *dataset Binary* nilai MCC dan G-mean yang paling tinggi menggunakan SMOTE-ENN + *Random Forest* dengan nilai tertinggi MCC = 0.99 dan nilai G-mean = 0.99 dari nilai MCC dan G-mean diatas akurasi sudah bagus karena nilai MCC dan G-mean yang mendekati 1 menunjukkan bahwa model memiliki performa klasifikasi yang sangat baik dalam menangani data tidak seimbang dengan menggunakan Metode SMOTE-Tomek + *Random Forest* dapat mencapai kinerja prediksi yang lebih besar untuk menangani *dataset* tidak seimbang *Binary*. Hal tersebut menunjukkan bahwa proses penanganan terhadap distribusi kelas yang tidak seimbang pada tahap *preprocessing* data memberikan pengaruh terhadap nilai akurasi MCC maupun G-mean metode SMOTE-Tomek + *Random Forest*.

**Kata kunci:** *binary*, data tidak seimbang, SMOTE-Tomek, SMOTE-ENN

## 1. Pendahuluan

Perkembangan teknologi informasi dan data analisis saat ini mendorong meningkatnya pemanfaatan *machine learning* dalam berbagai bidang. Namun, salah satu permasalahan umum yang sering muncul dalam penerapan *machine learning* adalah kondisi data tidak seimbang (*imbalanced data*) yaitu ketidakseimbangan jumlah data antar kelas, di mana kelas mayoritas jauh lebih dominan dibandingkan kelas minoritas (Aprianti et al., 2025). Kondisi ini dapat menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas sehingga menurunkan kemampuan model dalam mendeteksi kelas minoritas yang sering kali justru lebih penting (Yulian Pamuji et al., 2024).

Salah satu pendekatan yang banyak digunakan untuk mengatasi permasalahan data tidak seimbang (*imbalanced data*) adalah *resampling*, khususnya metode *oversampling*. *Synthetic Minority Over-sampling Technique (SMOTE)* merupakan metode yang populer karena mampu menghasilkan data sintesis pada kelas minoritas tanpa sekadar menduplikasi data yang ada. Meskipun demikian,

SMOTE memiliki kelemahan berupa potensi munculnya *overlapping* antar kelas dan peningkatan *noise* pada data, yang dapat menurunkan kualitas model klasifikasi (Latief et al., 2024).

Untuk mengatasi kelemahan tersebut, dikembangkan metode *hybrid* yang menggabungkan SMOTE dengan teknik *undersampling*, di antaranya SMOTE-Tomek dan SMOTE-ENN. SMOTE-Tomek menggabungkan SMOTE dengan Tomek Links untuk menghapus pasangan data yang ambigu di sekitar batas kelas, sehingga dapat memperjelas pemisahan antar kelas (Assyifa & Luthfiarta, 2024). Sementara itu, SMOTE-ENN menggabungkan SMOTE dengan *Edited Nearest Neighbor (ENN)* yang berfungsi menghapus data yang berpotensi sebagai *noise* berdasarkan kedekatan tetangga terdekatnya (Hermaliani & Ernawati, 2024).

Metode SMOTE-Tomek dan SMOTE-ENN merupakan metode untuk mengatasi permasalahan *dataset* tidak seimbang *Binary*, SMOTE-Tomek merupakan metode *hybrid resampling* yang menggabungkan SMOTE dengan teknik *undersampling* Tomek Links untuk menangani

permasalahan data tidak seimbang (*imbalanced data*). SMOTE-ENN merupakan metode *hybrid resampling* yang menggabungkan SMOTE dengan teknik undersampling Edited Nearest Neighbor (ENN) untuk mengurangi *noise* pada data (Misdran et al., 2023).

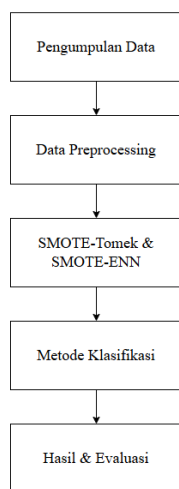
Untuk mengatasi permasalahan *dataset* tidak seimbang adalah dengan menyeimbangkan distribusi kelas tidak seragam di antara kelas-kelas dengan komparasi menggunakan metode SMOTE-Tomek dan SMOTE-ENN supaya jumlahnya seimbang dari kelas mayoritas (negatif) maupun kelas minoritas (positif) (Damari et al., 2025). Metode klasifikasi yang digunakan untuk mencari nilai akurasi yang menggunakan *machine learning*, kemudian *dataset* tersebut diuji dan di evaluasi yang akan dipilih metode yang menghasilkan nilai akurasi paling tinggi (Pamuji et al., 2021).

Dari uraian permasalahan penelitian ini adalah bagaimana menangani data tidak seimbang *Binary* pada data publik dari kumpulan data repositori KEEL dengan komparasi metode SMOTE-Tomek dan SMOTE-ENN dengan menggunakan metode klasifikasi. Penelitian ini bertujuan untuk mempertahankan jumlah kelas yang tidak seimbang agar tetap ideal dengan melakukan imputasi pada *dataset* yang tidak seimbang *Binary* dengan menguji metode SMOTE-Tomek dan SMOTE-ENN (Pamuji & Ramadhan, 2021).

## 2. Metode

### 2.1 Alur Penelitian

Pada penelitian ini, data yang digunakan adalah data tidak seimbang *Binary* dari kumpulan repositori KEEL. Data tidak seimbang *Binary* ini tersebut akan diolah menggunakan metode SMOTE-Tomek dan SMOTE-ENN. Dalam penelitian ini akan dilakukan beberapa tahap seperti yang ada pada Gambar 1.



Gambar 1. Alur Penelitian

### 2.2 Pengumpulan Data

Teknik pengumpulan data publik yang dilakukan dengan mempersiapkan data tidak seimbang dari kumpulan repositori KEEL <https://sci2s.ugr.es/keel/imbalanced.php> dengan berjumlah 5 *dataset Binary* yang terdiri dari 2 *class*, kemudian IR merupakan jumlah dari *imbalance ratio* tiap masing-masing data tidak seimbang, *instance* merupakan jumlah keseluruhan data tidak seimbang dan *atribut* merupakan jumlah atribut dari data tidak seimbang (Amin & Dinika, n.d.). Penelitian ini menggunakan beberapa *dataset binary imbalanced* dari repositori KEEL dengan tingkat *imbalance ratio* yang berbeda-beda untuk meningkatkan validitas dan generalisasi hasil penelitian. Data yang telah dikumpulkan dari data publik pada Tabel 1.

Tabel 1. *Dataset* Data Tidak Seimbang *Binary*

<i>Dataset</i>	IR ( <i>Imbalance Ratio</i> )	<i>Instance</i>	<i>Atribut</i>
wisconsin	1.86	683	9
page-blocks0	8.79	5472	10
ecoli3	8.6	336	7
segment0	6.02	2308	19

Dari Tabel 1 yang menunjukkan *dataset* yang terdiri dari 2 *class* di setiap masing-masing *dataset* dan IR (*Imbalance Ratio*) berbeda-beda tergantung tingkat tidak seimbangan dari *dataset*nya.

### 2.3 Data preprocessing

*Data transformation* untuk mengubah *dataset* dalam bentuk yang sesuai dalam proses data mining. *Normalization* dilakukan untuk menskalakan nilai *class* dalam rentang *class* dalam rentang nilai 1 dan 2 untuk *dataset binary*. *Dataset* terlebih dahulu dibagi menjadi data training sebesar 80% dan data testing sebesar 20%. Proses *resampling* menggunakan metode SMOTE-Tomek dan SMOTE-ENN hanya diterapkan pada data training untuk menghindari terjadinya *data leakage* (Aprianti et al., 2025).

### 2.4 SMOTE-Tomek

SMOTE-Tomek merupakan metode *hybrid resampling* yang menggabungkan teknik *oversampling* menggunakan Synthetic Minority Over-sampling Technique (*SMOTE*) dan undersampling menggunakan Tomek Links. *SMOTE* bekerja dengan membangkitkan data sintesis pada kelas minoritas berdasarkan interpolasi antara sampel minoritas dan tetangga terdekatnya (*K-Nearest Neighbors*) (Qolbu et al., 2025). Setelah proses *SMOTE*, metode Tomek Links diterapkan untuk menghapus pasangan data dari dua kelas berbeda yang saling berdekatan dan berpotensi berada pada area *overlap*. Dengan menggabungkan kedua teknik tersebut, SMOTE-Tomek tidak hanya

menyeimbangkan distribusi kelas, tetapi juga mengurangi *noise* dan *overlap* antar kelas, sehingga dapat meningkatkan performa model klasifikasi pada data tidak seimbang (Sun et al., 2023).

## 2.5 SMOTE-ENN

SMOTE-ENN merupakan metode *hybrid resampling* yang mengombinasikan SMOTE dengan Edited Nearest Neighbors (ENN). Tahap pertama dilakukan dengan SMOTE untuk menambah jumlah sampel pada kelas minoritas melalui pembentukan data sintesis (Eltehewy et al., 2023). Tahap selanjutnya adalah penerapan ENN sebagai metode pembersihan data (*data cleaning*). ENN menghapus sampel yang label kelasnya berbeda dengan mayoritas tetangga terdekatnya. Proses ini dapat menghapus data mayoritas maupun minoritas yang dianggap sebagai *noise* atau *outlier* (Jantung, 2023).

## 2.6 Metode Klasifikasi

Logistic Regression merupakan metode klasifikasi statistik yang digunakan untuk memodelkan hubungan antara variabel independen dan variabel dependen biner. Model Logistic Regression menentukan parameter dengan memaksimalkan fungsi *likelihood* menggunakan pendekatan optimasi (Geologi & Pertambangan, 2025). Klasifikasi dilakukan dengan menetapkan ambang batas (*threshold*) probabilitas tertentu untuk menentukan kelas data (Nugroho & Harini, 2024). Dalam fungsi logistik, variabel dependen adalah variabel biner yang berisi data berkode 1 (berhasil) atau 0 (gagal). Metode Logistic Regression menggunakan parameter *solver = lbfgs* dan *max\_liter = 1000*. Persamaan (1) menunjukkan perhitungan fungsi Logistic Regression yang digunakan.

$$p_i = \frac{1}{1 + e^{-(w^T x_i + b)}} \quad (1)$$

K-Nearest Neighbor (KNN) merupakan metode klasifikasi berbasis *instance* yang menentukan kelas suatu data berdasarkan mayoritas kelas dari k tetangga terdekatnya. Jarak antar data biasanya dihitung menggunakan metrik jarak, seperti Euclidean Distance (Agustian & Ramadhani, 2022). Pada proses klasifikasi, KNN tidak membangun model secara eksplisit, melainkan menyimpan seluruh data latih dan melakukan perhitungan jarak saat proses prediksi. Oleh karena itu, kualitas distribusi data latih sangat berpengaruh terhadap performa KNN (Illawati et al., 2025). Metode KNN menggunakan parameter jarak  $k=10$ . Persamaan (2) menunjukkan perhitungan fungsi KNN yang digunakan.

$$dis(x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \quad (2)$$

Random Forest merupakan metode klasifikasi berbasis *ensemble learning* yang menggabungkan banyak *decision tree* untuk meningkatkan akurasi dan stabilitas model. Pada setiap node dalam pohon keputusan, Random Forest memilih subset fitur secara acak untuk menentukan pemisahan terbaik. Strategi ini bertujuan untuk mengurangi korelasi antar pohon sehingga mampu menekan risiko *overfitting* yang sering terjadi pada *decision tree* Tunggal (Aryabima et al., 2025). Persamaan (3) dan (4) menunjukkan perhitungan fungsi Random Forest yang digunakan. Random Forest menggunakan 100 *decision tree* dengan parameter *random\_state = 42* untuk menjaga konsistensi hasil eksperimen.

$$Entropy(S) = \sum_{i=1}^n -P_i \log_2 P_i \quad (3)$$

$$Gain(S, f) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (4)$$

## 2.7 Hasil dan Evaluasi

Pada tahapan ini dilakukan evaluasi terhadap tingkat akurasi dari masing-masing metode untuk melihat kinerja setiap metode yang digunakan. Pada penelitian ini metode SMOTE-Tomek dan SMOTE-ENN akan dievaluasi dan divalidasi menggunakan alat ukur Confusion Matrix, MCC dan G-mean.

MCC sebagai memperhitungkan positif maupun negatif dan umumnya dianggap sebagai ukuran seimbang yang dapat digunakan bahkan jika kelas memiliki ukuran yang sangat berbeda. Persamaan (5) menunjukkan perhitungan fungsi MCC yang digunakan.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

G-mean sebagai nilai rata-rata yang diperoleh dengan mengalikan semua data dalam suatu kelompok sampel. Persamaan (6) menunjukkan perhitungan fungsi G-mean yang digunakan.

$$G_{mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (6)$$

## 3. Hasil dan Pembahasan

Tahapan proses data tidak seimbang menggunakan metode klasifikasi *dataset Binary* yang pertama proses *data preprocessing* dengan *normalization* dilakukan untuk menskalakan nilai class dalam rentang nilai 1 dan 2 untuk *dataset Binary*, kemudian *dataset Binary* akan di proses menggunakan metode klasifikasi yang digunakan antara lain Logistic Regression, KNN dan *Random Forest*. Dari hasil metode klasifikasi *dataset Binary* tersebut akan diambil nilai MCC dan G-mean untuk di analisa nantinya dengan hasil nilai MCC dan G-mean *dataset Binary* menggunakan SMOTE-Tomek dan SMOTE-ENN.

**Binary SMOTE-Tomek dan SMOTE-ENN (wisconsin)**

Evaluasi pada penelitian ini dari hasil nilai MCC dan G-mean dari *dataset Binary* (wisconsin) menggunakan SMOTE-Tomek dan SMOTE-ENN pada Tabel 2.

Tabel 2. *Binary* SMOTE-Tomek & SMOTE-ENN

Metode	SMOTE-Tomek		SMOTE-ENN	
	MCC	G-mean	MCC	G-mean
Logistic Regression	0.95	0.97	0.94	0.97
KNN (10)	0.93	0.96	0.95	0.97
<b>Random Forest</b>	<b>0.96</b>	<b>0.98</b>	<b>0.95</b>	<b>0.98</b>

Random Forest menghasilkan performa terbaik karena metode ini mampu menangani data hasil resampling yang memiliki distribusi kompleks dan non-linear. Selain itu, Random Forest memiliki kemampuan mengurangi *overfitting* melalui mekanisme *ensemble learning*.

Dari Tabel 2 untuk hasil tertinggi dari nilai MCC dan G-mean pada *dataset Binary* (wisconsin) menggunakan SMOTE-Tomek adalah metode klasifikasi Random Forest dengan nilai tertinggi MCC = 0.96 dan nilai G-mean = 0.98 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode SMOTE-ENN dengan metode klasifikasi Random Forest. nilai tertinggi MCC = 0.95 dan nilai G-mean = 0.98 dari metode yang lain seperti Logistic Regression dan KNN.

**Binary SMOTE-Tomek dan SMOTE-ENN (page-blocks0)**

Evaluasi pada penelitian ini dari hasil nilai MCC dan G-mean dari *dataset Binary* (page-blocks0) menggunakan SMOTE-Tomek dan SMOTE-ENN pada Tabel 3.

Tabel 3. *Binary* SMOTE-Tomek & SMOTE-ENN

Metode	SMOTE-Tomek		SMOTE-ENN	
	MCC	G-mean	MCC	G-mean
Logistic Regression	0.70	0.93	0.71	0.93
KNN (10)	0.79	0.95	0.77	0.95
<b>Random Forest</b>	<b>0.87</b>	<b>0.96</b>	<b>0.85</b>	<b>0.97</b>

Random Forest menghasilkan performa terbaik karena metode ini mampu menangani data hasil resampling yang memiliki distribusi kompleks dan non-linear. Selain itu, Random Forest memiliki kemampuan mengurangi *overfitting* melalui mekanisme *ensemble learning*.

Dari Tabel 3 untuk hasil tertinggi dari nilai MCC dan G-mean pada *dataset Binary* (page-blocks0) menggunakan SMOTE-Tomek adalah metode Klasifikasi Random Forest dengan nilai tertinggi

MCC = 0.87 dan nilai G-mean = 0.96 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode SMOTE-ENN dengan metode Klasifikasi Random Forest. nilai tertinggi MCC = 0,85 dan nilai G-mean = 0,97 dari metode yang lain seperti Logistic Regression dan KNN.

**Binary SMOTE-Tomek dan SMOTE-ENN (ecoli3)**

Evaluasi pada penelitian ini dari hasil nilai MCC dan G-mean dari *dataset Binary* (ecoli3) menggunakan SMOTE-Tomek dan SMOTE-ENN pada Tabel 4.

Tabel 4. *Binary* SMOTE-Tomek & SMOTE-ENN

Metode	SMOTE-Tomek		SMOTE-ENN	
	MCC	G-mean	MCC	G-mean
Logistic Regression	0.57	0.87	0.54	0.86
KNN (10)	0.54	0.80	0.47	0.83
<b>Random Forest</b>	<b>0.84</b>	<b>0.91</b>	<b>0.69</b>	<b>0.94</b>

Random Forest menghasilkan performa terbaik karena metode ini mampu menangani data hasil resampling yang memiliki distribusi kompleks dan non-linear. Selain itu, Random Forest memiliki kemampuan mengurangi *overfitting* melalui mekanisme *ensemble learning*.

Dari Tabel 4 untuk hasil tertinggi dari nilai MCC dan G-mean pada *dataset Binary* (ecoli3) menggunakan SMOTE-Tomek adalah metode Klasifikasi Random Forest dengan nilai tertinggi MCC = 0.84 dan nilai G-mean = 0.91 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode SMOTE-ENN dengan metode Klasifikasi Random Forest. nilai tertinggi MCC = 0,69 dan nilai G-mean = 0,94 dari metode yang lain seperti Logistic Regression dan KNN.

**Binary SMOTE-Tomek dan SMOTE-ENN (segment0)**

Evaluasi pada penelitian ini dari hasil nilai MCC dan G-mean dari *dataset Binary* (segment0) menggunakan SMOTE-Tomek dan SMOTE-ENN pada Tabel 5.

Tabel 5. *Binary* SMOTE-Tomek & SMOTE-ENN

Metode	SMOTE-Tomek		SMOTE-ENN	
	MCC	G-mean	MCC	G-mean
Logistic Regression	0.98	0.99	0.98	0.98
KNN (10)	0.89	0.98	0.87	0.97
<b>Random Forest</b>	<b>0.99</b>	<b>0.99</b>	<b>0.98</b>	<b>0.99</b>

Random Forest menghasilkan performa terbaik karena metode ini mampu menangani data hasil resampling yang memiliki distribusi kompleks dan

non-linear. Selain itu, Random Forest memiliki kemampuan mengurangi *overfitting* melalui mekanisme *ensemble learning*.

Dari Tabel 5 untuk hasil tertinggi dari nilai MCC dan G-mean pada *dataset Binary* (segment0) menggunakan SMOTE-Tomek adalah metode Klasifikasi Random Forest dengan nilai tertinggi MCC = 0,99 dan nilai G-mean = 0,99 dari metode yang lain seperti Logistic Regression dan KNN. Kemudian menggunakan metode SMOTE-ENN dengan metode Klasifikasi Random Forest. nilai tertinggi MCC = 0,98 dan nilai G-mean = 0,99 dari metode yang lain seperti Logistic Regression dan KNN.

#### 4. Kesimpulan

Berdasarkan hasil eksperimen yang telah dilakukan dari penelitian ini yaitu bahwa pengujian untuk *dataset Binary* nilai MCC dan G-mean yang paling tinggi menggunakan SMOTE-Tomek + Random Forest dengan nilai tertinggi MCC = 0,99 dan nilai G-mean = 0,99 dari nilai MCC dan G-mean diatas akurasi sudah bagus karena nilai MCC dan G-mean yang mendekati 1 menunjukkan bahwa model memiliki performa klasifikasi yang sangat baik dalam menangani data tidak seimbang dengan menggunakan Metode SMOTE-Tomek + Random Forest dapat mencapai kinerja prediksi yang lebih besar untuk menangani *dataset* tidak seimbang *Binary*. Metode SMOTE-Tomek yang dikombinasikan dengan Random Forest memberikan performa terbaik pada dataset binary imbalanced berdasarkan metrik MCC dan G-mean. Metode ini mampu meningkatkan kualitas klasifikasi dibandingkan SMOTE-ENN. Namun, penelitian selanjutnya disarankan untuk mengembangkan dan membandingkan metode *hybrid resampling* lainnya yang mengombinasikan teknik *oversampling* dan *undersampling*, diharapkan dapat memberikan gambaran yang lebih komprehensif mengenai efektivitas metode dalam menangani permasalahan data tidak seimbang.

#### Daftar Pustaka:

- Agustian, S., & Ramadhani, S. (2022). *Jurnal Computer Science and Information Technology (CoSciTech) menggunakan algoritma lexicrank*. 3(3), 371–381.
- Amin, F. M., & Dinika, Q. A. (n.d.). *Penerapan Algoritma Hybrid Sampling SMOTE-TomekLink dan Random Forest untuk Klasifikasi Penyakit Diabetes*. September 2025, 24–31.
- Aprianti, S. D., Alzami, F., Rizqa, I., Pramunendar, R. A., & Megantara, R. A. (2025). *Comparative Analysis of PCOS Classification Using Random Forest: Integration of Mutual Information, SMOTE-Tomek, and Outlier Handling*. 10(1), 78–87.
- Aryabima, M. I., Roeswidiah, R., & Pudoli, A. (2025). *Deteksi Dini Penyakit Stroke pada Data Tidak Seimbang Menggunakan SMOTE dan Random*. 13, 141–146.
- Assyifa, D. S., & Luthfiarta, A. (2024). *SMOTE-Tomek Re-sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-class Classification*. 9(2), 151–160.
- Damari, A., Azhima, T., Siswa, Y., Pranoto, W. J., Muhammadiyah, U., Timur, K., & Timur, K. (2025). *IMPLEMENTATION OF THE PSO-SMOTE METHOD ON THE NAIVE BAYES ALGORITHM TO ADDRESS CLASS IMBALANCE IN LANDSLIDE DISASTER DATA PENERAPAN METODE PSO-SMOTE PADA ALGORITMA NAIVE BAYES UNTUK MENGATASI CLASS IMBALANCE*. 10(1), 332–343.
- Eltehewy, R., Abouelfarag, A., & Saleh, S. N. (2023). *Efficient Classification of Imbalanced Natural Disasters Data Using Generative Adversarial Networks for Data Augmentation*. *ISPRS International Journal of Geo-Information*, 12(6), 245. <https://doi.org/10.3390/ijgi12060245>
- Geologi, P., & Pertambangan, D. A. N. (2025). *PENERAPAN NAIVE BAYES DENGAN BALANCING DATA SMOTE UNTUK*. 2(1), 1–8.
- Hermaliani, E. H., & Ernawati, M. (2024). *Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian*. 4(1).
- Illawati, A. R., Hadiana, A. I., Informatika, J. T., Jenderal, U., Yani, A., Artikel, I., & Imbalance, D. (2025). *Klasifikasi Penyakit Monkeypox dengan XGBoost dan SMOTE untuk Penanganan Data Tidak Seimbang*. 35–44. <https://doi.org/10.33364/algoritma/v.22-2.2349>
- Jantung, G. (2023). *Perbandingan model klasifikasi random forest dengan resampling dan tanpa resampling pada pasien penderita gagal jantung I*. 12, 136–145. <https://doi.org/10.14710/j.gauss.12.1.136-145>
- Latief, M. A., Nabila, L. R., Miftakhurrahman, W., Ma, S., Tantyoko, H., & Latief, M. A. (2024). *Handling Imbalance Data Using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification Corresponding Author*: 3(1), 11–18. <https://doi.org/10.30812/IJECSA.v3i1.3758>
- Misdrum, M., Muljono, Purwanto, & Noersasongko, E. (2023). *Gradually Generative Adversarial Networks Method for Imbalanced Datasets*. *International Journal of Advanced Computer Science and Applications*, 14(4), 51–58. <https://doi.org/10.14569/IJACSA.2023.0140408>
- Nugroho, A., & Harini, D. (2024). *Teknik Random Forest untuk Meningkatkan Akurasi Data Tidak*

- Seimbang*. 2(2), 128–140.
- Pamuji, F. Y., Dwi, S., & Putri, A. (2021). *Komparasi Metode SMOTE dan ADASYN Untuk Penanganan Data Tidak Seimbang MultiClass*. 331–338.
- Pamuji, F. Y., & Ramadhan, V. P. (2021). *Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotherapy*. *Jurnal Teknologi Dan Manajemen Informatika*, 7(1), 46–50. <https://doi.org/10.26905/jtmi.v7i1.5982>
- Qolbu, A. A., Fitriyati, N., & Inayah, N. (2025). *Performa Naïve Bayes, SVM, dan IndoBERT pada Analisis Sentimen Twitter IndiHome dengan Strategi Penanganan Data Tidak Seimbang*. 814(1), 29–44. <https://doi.org/10.14421/fourier.2025.141.29-44>
- Sun, Y., Lee, J., Kim, S., Seon, J., Lee, S., Kyeong, C., & Kim, J. (2023). *Energy Theft Detection Model Based on VAE-GAN for Imbalanced Dataset*. *Energies*, 16(3). <https://doi.org/10.3390/en16031109>
- Yulian Pamuji, F., Ahmad Rofiqul Muslikh, Rizza Muhammad Arief, & Delviana Muti. (2024). *Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil*. *Jurnal Informatika Polinema*, 10(2), 257–264. <https://doi.org/10.33795/jip.v10i2.5031>